
UNIVERSIDAD CARLOS III DE MADRID

# Sistema de síntesis de texto a voz femenina en español con control prosódico basado en Mbrola

**Autora: Emilia Méndez Barrios**

**Tutor: Javier Fernández de Gorostiza**

**Codirector: Fernando Alonso Martín**

# Agradecimientos

Después de terminar de escribir el documento tengo bastantes más ganas de agradecerle esto a todo el mundo dando un beso y un abrazo, pero eso no llega a la posteridad así que lo tendré que hacer aquí.

A los que me habeis ayudando, revisando...y revisando...y revisando este documento hasta que os han sangrado los ojos, ayudandome con los formularios y orientandome con vuestro proyecto.

A los compañeros del laboratorio, que me han echado una mano cuando lo necesitaba.

A mis padres, que son las personas gracias a las que estoy aquí, gracias a las que soy quién soy y que me han aguantado mis momentos de estrés. Quienes me han dado la posibilidad de dedicarme a lo que me gusta, pareciendo siempre que no les costaba ningún trabajo. Y porque tengo la certeza de que pase lo que pase siempre van a estar ahí.

A mi hermana, que es la mejor hermana del mundo, y que es mi confidente y mi consejera. Aun no lo entiendo pero dentro de seis años lo entenderé.

Al resto de mi familia por poder crecer al lado de tanta gente que me quiere.

A Elena, que todos estos años me ha amenizado las temporadas de exámenes a base de llamadas telefónicas, que conseguían que estudiara mucho más rápido, y menos. La amistad que yo doy por hecha, mucha gente se morirá sin conocerla.

A Kike, por intentar que me echaran de la carrera, y por todo un tiempo en el que me ha llenado de vida y me ha hecho verme de otra manera.

Y ahora me debería poner a enumerar amigos, a veces dicen que se pueden contar con los dedos de una mano, pues a mi me faltan. Gracias a todos los que estais ahí día a día. Es genial sentirse siempre tan querida y apoyada.

A mis compañeros de universidad, por hacer de estos años mucho más fáciles. Conseguir que quiera venir a la universidad solo por comer con vosotros. Bueno por intentar comer con vosotros.

A todos mis compañeros de otras actividades, que completan mi existencia, y mi tiempo, haciéndome disfrutar de su compañía y permitiéndome conocer mucho más mundo con sus vivencias.

# Resumen

En los robots sociales buscamos que un humano y una máquina puedan comunicarse de la manera más natural posible, persiguiendo una comunicación completa.

El presente proyecto busca incorporar en el habla del robot una variante emocional que pueda dotar al mismo mensaje de más información.

Se ha diseñado de manera que las emociones se obtengan a partir de un espacio emocional continuo, con el que podemos conseguir un gran número de emociones a partir de tres parámentros (Arousal, Valence, Power).

PALABRAS CLAVE: Robot social, comunicación, emoción, Mbrola, comunicación en tiempo real, Interacción Humano-Máquina, prosodia, entonación, síntesis de voz.

# Índice general

# Índice de figuras

# Capítulo 1

# Introducción

## Motivación del proyecto y objetivos

El presente proyecto pretende mejorar la comunicación entre un robot social y el usuario, enfatizando la comunicación oral del sistema mediante la expresión de emociones. El programa propuesto pretende funcionar de forma coordinada con un hipotético motor emotivo, encargado de la génesis de las emociones del robot, de forma que el mensaje oral quede reforzado mediante entonaciones coherentes con el contexto emocional del robot.

Un robot social es aquel que interactúa y se comunica con las personas (de forma sencilla y agradable) siguiendo comportamientos, patrones y normas sociales. Para eso se necesita que disponga de habilidades que se ubican dentro del dominio de la llamada inteligencia social. Encontramos la mayor parte de los robots sociales dentro del entorno universitario, pero la investigación tiende, poco a poco, a hacerlos participes en la vida cotidiana como robots de compañía. Sirviendo para el apoyo en el cuidado y acompañamiento de personas mayores, discapacitadas y niños.

Si nos centramos en la comunicación con el robot, la mayor parte de esta se realizara de manera hablada, y la comunicación debe ser lo más precisa y adaptada posible. La mayor parte de los robots o las voces robóticas se expresan de manera neutra o son capaces de reproducir las emociones básicas, sin embargo, la comunicación será más rica si podemos generar un número indefinido de emociones. Esta nueva expresividad hara más efectiva la comunicación en momentos en el que los matices en el tono abren el camino a los elementos verbales y por tanto a la consecución de un objetivo.

La comunicación oral humana comprende tanto elementos verbales (palabras empleadas) como elementos no verbales (entonación, ritmo, actitud corporal, etcétera). Son precisamente estos elementos no verbales los que el cerebro procesa a mayor velocidad, es decir, recibimos la información transmitida a través del tono antes que aquella puramente

contenida en las palabras del hablante. Éste fenómeno nos lleva a tomar decisiones conductuales al margen del mensaje verbal, y por tanto resulta fundamental proveer a un robot social de una prosodia[1] adecuada al mensaje con objeto de proporcionar a la comunicación su vertiente no verbal. De esta manera se consigue una comunicación más completa, enfatizando la información verbal, matizándola, ampliándola o incluso contradiciéndola.

El objetivo general de este proyecto es lograr que el mensaje verbal emitido por el robot sea acompañado de una entonación acorde con a la emoción generada por el motor emotivo a partir del contexto. A nivel técnico se plantean tres objetivos secundarios:

1. Obtener dicha entonación a tiempo real. Se ha utilizado para este efecto Mbrola por su pequeño tamaño de base de datos y por que su algoritmo genera poco coste computacional.

2. Lograr una entonación que refleje un espacio emocional continuo. Se ha utilizado un espacio basado en tres variables ($R^3$), acorde con los estudios de Schröder y Pereira.[26][21]

# Desarrollo y cronología

1. Creación de base de datos con difonos Mbrola.

   *a*) Estudio del español y de los difonos necesarios para la base de datos.

   *b*) Creación del corpus para grabación (texto que contiene todos los difonos del lenguaje).

   *c*) Grabación del corpus.

   *d*) Recorte de difonos

   *e*) Resintetización y creación de base de datos.

2. Creación de Text to Speech (TtS) para leer textos con Mbrola.

3. Implementación de "lecturas emotivas" para cinco emociones (alegría, calma, confort, tristeza y enfado)[2].

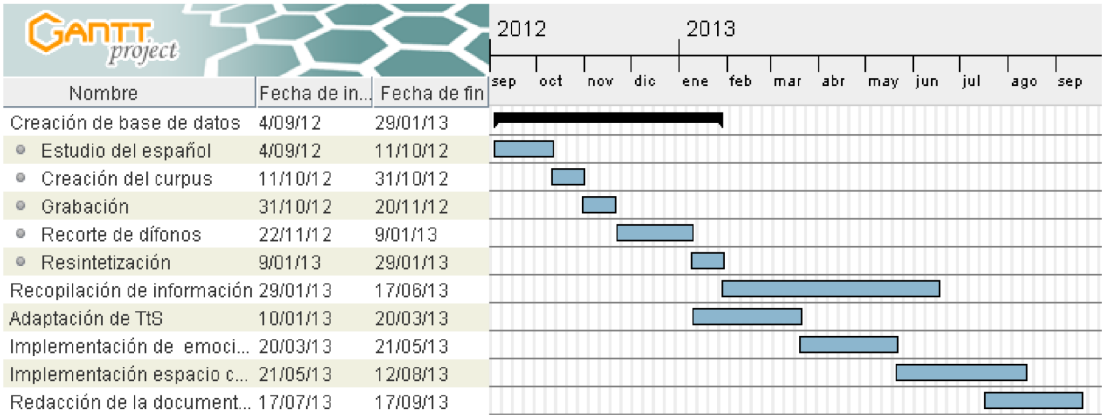
4. Relación del espacio de emociones con los parámetros paralingüísticos. Obtención del mensaje con sus características lingüísticas y para lingüísticas.

[1] Usaremos este término varias veces a lo largo del documento. La prosodia se ocupa de la entonación del mensaje y la acentuación.

[2] Esta implementación se hace como resultado intermedio (conseguir algunas emociones) antes de proseguir a la consecución del objetivo final (crear el campo de emociones continuo).

5. Redacción de la documentación.

El siguiente gráfico muestra el diagráma de Gantt del proyecto.




| Nombre | Fecha de in... | Fecha de fin |
|---|---|---|
| Creación de base de datos | 4/09/12 | 29/01/13 |
| Estudio del español | 4/09/12 | 11/10/12 |
| Creación del curpus | 11/10/12 | 31/10/12 |
| Grabación | 31/10/12 | 20/11/12 |
| Recorte de difonos | 22/11/12 | 9/01/13 |
| Resintetización | 9/01/13 | 29/01/13 |
| Recopilación de información | 29/01/13 | 17/06/13 |
| Adaptación de TtS | 10/01/13 | 20/03/13 |
| Implementación de emoci... | 20/03/13 | 21/05/13 |
| Implementación espacio c... | 21/05/13 | 12/08/13 |
| Redacción de la document... | 17/07/13 | 17/09/13 |

Figura 1.0.1: Digrama de Gantt

# Capítulo 2

# Base científica empleada

Numerosos estudios [15, 4, 12, 14]realizados sobre la influencia y el impacto de cada elemento comunicativo en el receptor han arrojado los siguientes resultados:

- El 55 % de la información que recibimos corresponde al lenguaje corporal (postura, expresión facial, movimiento, etcétera...)
- El 38 % corresponde a la entonación y otros atributos de la voz como el timbre.
- Tan solo un 7 % del mensaje corresponde a las palabras utilizadas.

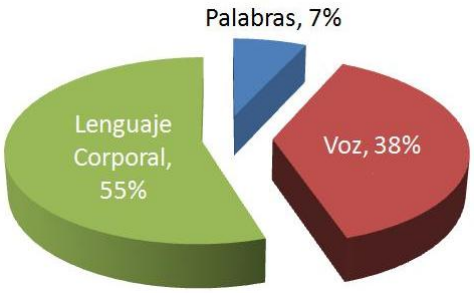



Figura 2.0.1: Recepción de la información en el acto comunicativo. Imagen procedente de : “http://ginaaran.wordpress.com”

Estos resultados ponen de manifiesto que sólo una pequeña parte del mensaje que recibimos se debe exclusivamente a la comunicación verbal oral, y que la mayor parte de lo procesado corresponde a factores no verbales. En otras palabras, la forma de expresarnos (entonación, características de la voz y lenguaje corporal) tiene mayor influencia que el contenido en sí.

## 2.1. La expresión de emociones

### La expresión no verbal

La siguiente figura resume los principales factores que forman parte de la comunicación no verbal:

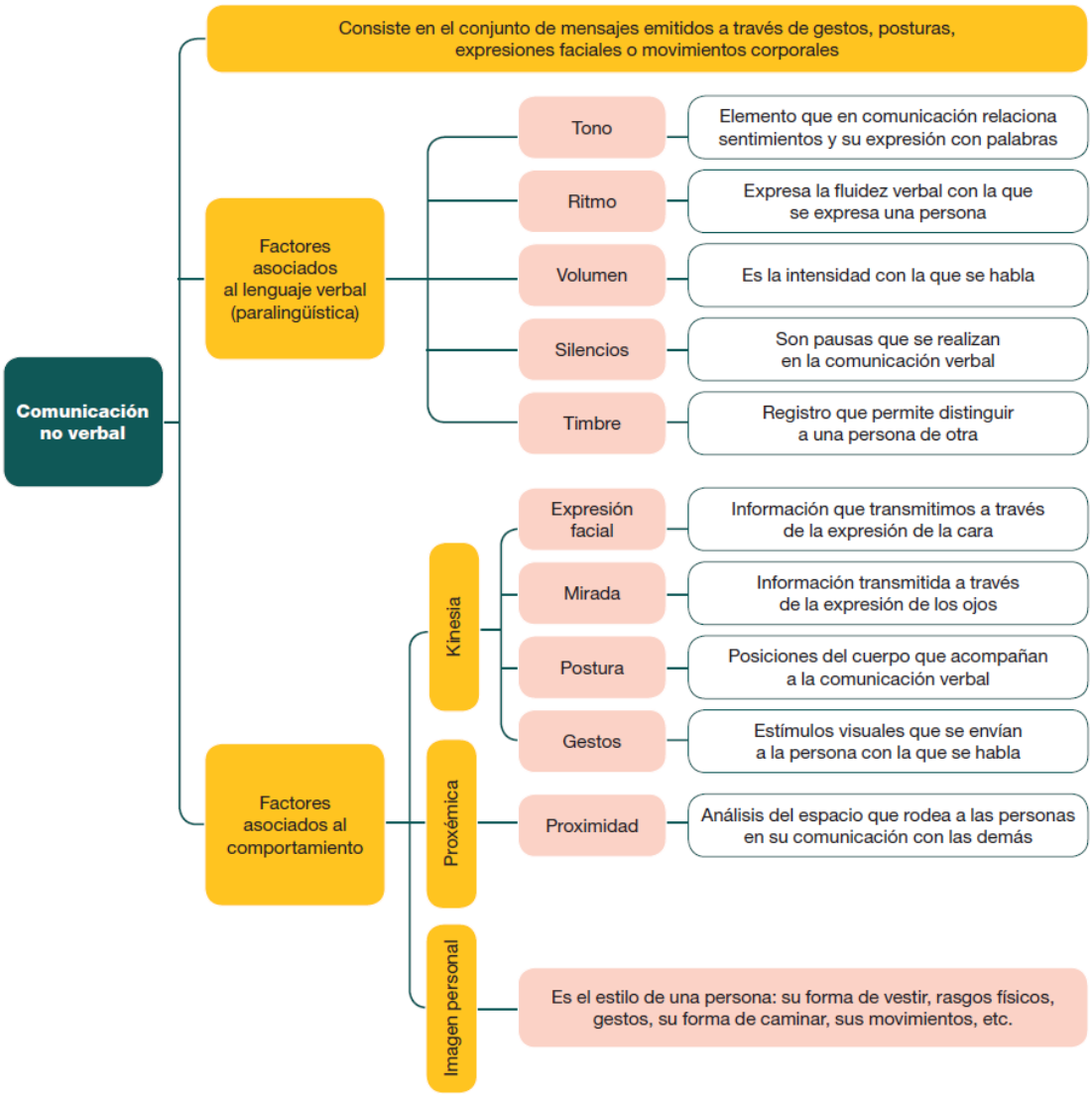



Figura 2.1.1: La comunicación no verbal([14])

A continuación se detallan cada uno de estos factores[6, 14], haciendo especial hincapié en aquellos que corresponden al ámbito de la paralingüística, ya que se encuentran especialmente relacionados con el proyecto propuesto.

### Kinestesia

Engloba los gestos y movimientos del cuerpo:

- Postura corporal. Una postura abierta indica una mayor disposición a interactuar, mientras que las posturas cerradas son menos invitadoras.
- Gestos. Movimientos del rostro y las manos con los que expresamos diversos estados de ánimo. Existen gestos aprendidos e innatos.
- Expresión facial. Regula la interacción y refuerza al receptor. Se ha determinado la existencia de seis expresiones faciales principales, las cuales expresan emociones genéricas: alegría, tristeza, asco, enfado, temor, e interés. Estas expresiones son reconocidas con mayor facilidad.
- Mirada. Es considerada uno de los mejores indicadores emocionales La dilatación de las pupilas indica interés y atracción, el número de veces que parpadeamos está relacionado con el grado de nerviosismo y una mirada sostenida es un indicador de interés, agrado o sinceridad.
- Sonrisa. Expresa simpatía, alegría o felicidad. Además de ser atractiva es una forma de relajar tensión. Sin embargo, existen sonrisas con significados diferentes.
- Movimientos corporales completos. Forma de caminar, de sentarse, balanceo, etcétera...

### Proxémica

Se refiere a la proximidad existente entre hablantes. Considera la burbuja espacial personal adecuándose al lugar y momento en que nos encontramos. Depende de 5 factores:

- Grado de intimidad
- Personalidad
- Marco cultural.
- Edad
- Naturaleza del encuentro.

## Paralingüística: La expresión no verbal a través de la voz

El primer autor que define el concepto paralingüística es HILL(1956), para quien se trata de "una parte de la actividad comunicativa que se encuentra fuera del área de la micro lingüística". Asimismo, en la misma época, otro autor, SMITH(1953) propone definir el término como "el conjunto de vocalizaciones y cualidades de la voz", y afirma que "el habla no tiene lugar en el vacío, sino que se rodea de movimientos corporales y vocalizaciones"

Según estas primeras definiciones, la paralingüística englobaría todos los elementos de la comunicación no verbal. Sin embargo, en los años 80 aparecen una serie de autores que acotan el campo paralingüístico hasta el concepto que manejamos hoy. Podemos dividir a estos autores más recientes en dos corrientes, en función de los elementos que consideran como parte del paralenguaje.

En primer lugar, existen una serie de autores que consideran como paralenguaje sólo los rasgos prosódicos que no afectan al significado de la frase, pero que revelan factores extralingüísticos, es decir, únicamente aquellos rasgos que son vocales pero no verbales. Entre estos autores destacan CRYSTAL(1963) y ARGYLE(1978).

A su vez, otros autores incluyen como paralenguaje todos aquellos sonidos que no se ajusten a la estructura fonética de la lengua. Se trata por tanto de una perspectiva mucho más amplia que la primera, ya que incluyue elementos como interjecciones u onomatopeyas. Los máximos representantes de esta línea de pensamiento son POYATOS (1993)y AUSTIN(2002).

A efectos prácticos, en este proyecto consideraremos paralingüística como el conjunto de elementos que acompañan a la comunicación (oral o escrita), y que constituyen señales e indicios que transmiten información adicional, matizan, reafirman, aclaran o sugieren interpretaciones particulares de la información propiamente lingüística. En términos sencillos, el paralenguaje se podría definir como "todo aquello que está más allá de las palabras".

Existen estudios[6] en que se caracterizan los elementos paralingüísticos propios de los distintos tipos de comunicación:

- Paralingüística oral: La expresión oral comprende gran cantidad de elementos no lingüísticos como el ritmo del discurso, el tono o el volumen de la voz. Estos elementos permiten que el oyente comprenda con mayor facilidad lo que se pretende transmitir, y por lo tanto, refuerzan el contenido del lenguaje verbal.
- Paralingüística escrita: En el acto comunicativo a menudo se emplea un soporte gráfico, que cuenta con elementos paralingüísticos diferentes a los que se aplican a

la lengua oral. Son, por ejemplo, la tipografía empleada, el tamaño de letra, el color, la elección de mayúsculas y minúsculas, y, recientemente, el cada vez más frecuente uso de emoticonos, especialmente en el ámbito de la mensajería instantánea.

## Elementos paralingüísticos

### Tono e inflexión

Tono e inflexión son dos parámetros íntimamente relacionados. Se puede entender como tono la propiedad de los sonidos que los caracteriza como agudos o graves, en función de su frecuencia. No obstante, existen definiciones más amplias, como la que brinda la RAE, que define tono como "inflexión de la voz y modo particular de decir algo, según la intención o el estado de ánimo de quien habla". La inflexión, por su parte, es la elevación o atenuación que se hace con la voz, así como el paso de un tono a otro.

El tono y la inflexión que se emplean al hablar permiten comunicar actitudes, intenciones y emociones del hablante. Por ejemplo, un elemento verbal como puede ser un simple "Oh" adquiere significados muy diferentes en función del tono empleado. Un "Oh" ascendente indica sorpresa o contrariedad, mientras que un "Oh" descendente se interpreta como desilusión. El tono con el que nos dicen algo como "buenos días" nos sirve de indicador sobre el estado emocional del emisor en ese momento, al margen del mensaje verbal. Sirve, por tanto, como regulador entre el sentimiento y la expresión, entre lo sentido y lo verbalizado. Por ello, además de buscar las palabras adecuadas, las personas adaptamos constantemente nuestro tono e inflexiones al mensaje que pretendemos transmitir.

### Volumen

El volumen de la voz permite transmitir emociones y enfatizar algunas partes del discurso. Cada situación implica un rango de volumen que consideramos apropiado, es decir, lo suficientemente elevado como para que todos puedan oír al hablante sin que resulte molesto. Un volumen excesivamente elevado suele ser síntoma de que el interlocutor pretende imponerse en la conversación, y está relacionado con la intención de mostrar autoridad, dominio, seguridad, o incluso puede ser indicador de alegría o enfado. Un volumen moderado transmite calma, y el volumen bajo se asocia a personas introvertidas o situaciones en que la conversación resulta inapropiada.

El contexto comunicativo (lugar, hora, interlocutores, situación, etcétera...) condiciona el rango de volumen que percibimos como adecuado.

### El ritmo

El ritmo hace referencia a la cantidad de palabras emitidas por unidad de tiempo (generalmente medido en palabras por minuto). El ritmo normal que garantiza que el mensaje sea inteligible es de entre cien y ciento cincuenta palabras por minuto. Si un emisor habla a ritmo muy lento, resulta aburrido, mientras que un ritmo demasiado ligero enerva, agobia y acelera al oyente.

### El timbre

La voz es como una huella dactilar, una característica personal e intransferible de cada individuo. El timbre es el conjunto de características de la voz que nos permite distinguir a unas personas de otras.

Los sonidos que escuchamos son complejos, es decir, están compuestos por varias ondas simultáneas que nosotros percibimos como un todo. El timbre depende de la cantidad de armónicos que componen el sonido de la voz, y de la intensidad de cada uno de ellos.[1]

### Fluidez verbal

Es la capacidad de expresar ideas con corrección y facilidad de forma espontánea. Un discurso desestructurado y poco fluido se percibe como confuso e indica un dominio pobre del lenguaje.

### Pausas y silencios

Las pausas funcionan, generalmente, como reguladores de cambio. Indican, por ejemplo, cambios de unidad gramatical, de tema o de turno de palabra. Existen dos tipos de pausas, reflexivas y fisiológicas. Son pausas reflexivas aquellas que resultan necesarias para comprender correctamente el mensaje verbal. Vienen determinadas por la estructura de la lengua. De esta manera, la expresión "Sandra entra" cambia claramente de significado dependiendo de si se efectúa una pausa entre las dos palabras o no. "Sandra, entra" implica una orden, mientras que "Sandra entra" informa de una acción. Las pausas fisiológicas son aquellas que resultan necesarias para el correcto funcionamiento del aparato respiratorio. En el caso de un robot, estas pausas se convierten evidentemente en un elemento innecesario.

Los silencios, lejos de implicar una ausencia de comunicación, constituyen un medio en ocasiones mucho más efectivo que los discursos o explicaciones verbales. Con el silencio,

[1] http://es.wikipedia.org/wiki/Timbre_%28ac%C3%BAstica%29

por ejemplo, se invita a hablar, a callar, se asiente o se muestra desacuerdo, entre muchas otras posibilidades.

### Claridad

Refleja la facilidad con la que un hablante es comprendido. Está relacionada con la velocidad, los defectos de pronunciación y de vocalización, los acentos locales, la capacidad de estructurar el mensaje verbal, etcétera...

### Tiempo de habla

Se refiere a la duración de las intervenciones de los interlocutores durante una conversación. Los tiempos de habla adecuados se perciben de formas muy diferentes dependiendo de la situación comunicativa. Por ejemplo, en una conversación entre dos amigos se percibiría como apropiado un equilibrio en los tiempos de habla de cada uno de los interlocutores, mientras que en el contexto de una conferencia esperamos que un solo interlocutor acapare la mayor parte del tiempo de habla.

Con el objeto de obtener una voz robótica capaz de comunicarse con el usuario con la máxima naturalidad, es necesario tener en cuenta cada uno de los aspectos anteriormente descritos, buscando imitar con la mayor precisión posible, los patrones que rigen la comunicación humana en cada uno de ellos.

# Capítulo 3

# Modelos de síntesis de voz

## 3.1. Síntesis de voz

Para sintetizar la voz humana existen varias técnicas, cada una de ellas tiene sus ventajas y sus desventajas, en síntesis de de voz la calidad se mide a partir de 2 factores:

**Naturalidad:** Describe en qué medida el sonido generado se asemeja al habla humana.

**Inteligibilidad:** Es la facilidad con la cual se entiende el significado del habla.

Estos dos factores son subjetivos, no se puede medir numéricamente más que por estudios estadísticos a los usuarios.

Las distintas tecnologías albergan distinta composición de estas cualidades. Describiremos a continuación esas tecnologías.

### Síntesis concatenativa

Se basa en la unión de segmentos de voz grabados. Esta síntesis produce resultados más naturales aunque la automatización de las uniones conlleva la perdida de la naturalidad.

Podemos diferenciar tres tipos[1]:

#### Síntesis específica de un dominio

La síntesis específica para un dominio concatena palabras y frases grabadas para crear salidas completas. Es la usada históricamente, en despertadores, relojes, etcétera...

[1] http://es.wikipedia.org/wiki/S%C3%ADntesis_de_habla

Se usa cuando hay poca variedad de oraciones y la prosodia y la entonación corresponde con las originales.

Tienen mucha naturalidad y inteligibilidad en el mensaje, pero podemos usarlo para un número determinado de mensajes concretos o necesitamos bases de datos muy grandes.

**Síntesis de difonos:**

Este tipo de síntesis usa una base de datos mínima que contiene todos los difonos existentes en el lenguaje, la base de datos contiene solo un ejemplo de cada difono.

El español contiene 800 difonos, y es uno de los idiomas que necesita bases de datos más pequeñas.

La prosodia de la oración se impone de mediante procesamiento digital de señales, por elemplo, PSOLA

Este tipo de tecnología obtiene, por norma general, una calidad de habla peor que la obtenida por selección de unidades, pero más natural de que se obtiene por síntesis de formantes.

Este tipo de voz suena robótica pero tiene como ventaja el pequeño tamaño de la base de datos y la cantidad de implementaciones libres, lo que la hace atractiva para la investigación.

**Síntesis por selección de unidades:**

Como se explica en [7]las frases a grabar deben contener múltiples instancias de cada difono y seleccionar la que tenga las características prosódicas más cercanas a las deseadas. Las distintas grabaciones deben respetar la distribución de frecuencias del lenguaje y el número de instancias de cada difono depende de cuánto se use en el idioma, por ejemplo, en español, muchas instancias de /la/ ; pocas de /pt/ . Esto conlleva varias horas de grabación.

La segmentación de los fonemas es semi-automática, es decir, la alineación esta forzada por sistemas de reconocimiento pero necesita corrección manual.

En la síntesis se busca encontrar en la base la secuencia que mejor cumpla la especificación dada. Esto depende de dos factores:

- (T) Cuánto respetan los difonos las características específicas (prosodia, contexto...)
- (J) Cómo de bien se concatenan (perceptiblemente) los difonos.

$$\hat{U} = \arg\min_{u} \sum_{t+1}^{T} T(S_t, U_t) + \sum_{t=1}^{T-1} (J(U_t, U_{t+1})$$

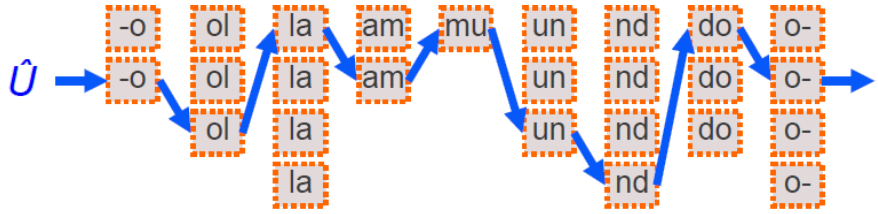



Figura 3.1.1: Síntesis por selección de formantes(A.Gravano)

Los resultados de este tipo de síntesis son más naturales que con otras técnicas.

Como contra, tenemos bases de datos muy grandes, del orden de GB. La búsqueda de difonos requiere un alto coste computacional, ya que las bases de datos son muy grandes aunque se usen técnicas de optimización como el clustering de difonos.

La calidad del audio puede ser muy mala si no hay buenos candidatos para las características deseadas.

## Síntesis de formantes

La síntesis de formantes crea la salida usando un modelo acústico, crea la onda de voz de manera artificial.

Estos sistemas crean una voz robótica, con poco naturalidad, pero a cambio, es muy inteligible, evita los defectos acústicos producidos por la frecuencia en los sistemas concatenativos Además, los programas suelen ser más pequeños ya que no necesitan base de datos de muestras de voz grabada.

## Síntesis articulatoria

Se basa en modelos computacionales del tracto vocal y el proceso de articulación. Pocos de esos modelos son suficientemente avanzados o eficientes computacionalmente para ser usados en sistemas comerciales de síntesis de voz.

## Otros

Existen otros modelos de sistemas como la síntesis híbrida que utiliza aspectos de la concatenativa y la de formantes para minimizar los defectos de la concatenación o la

Síntesis basada en HMM (modelos ocultos de Márkov).[2]

## 3.2. Variables acústicas relevantes para la síntesis del habla

El modelo acústico se representa como un grupo de parámetros que varían la forma de "sonar" de la frase.

Cada parámetro varía independientemente. Estos parámetros son los que se relacionan con cada eje de la emoción.

En el estudio [3] se separan estos parámetros en 4 categorías.

#### Parámetros de frecuencia

- Forma del acento: Describe la forma en la que varia la frecuencia en cada acento en la frase la pendiente o la rugosidad del la frecuencia entorno al acento.
- Media de frecuencia: Describe la media de un hablante normal.
- Pendiente del contorno: Describe la tendencia general del rango de la frecuencia en la expresión.
- Final de la expresión: Describe como cambia la frecuencia al final de la frase.
- Rango de frecuencia: Describe la banda de rangos.
- Línea de referencia: Línea que genera la entonación (enunciativa, interrogativa o exclamativa).

#### Parámetros de tiempo

- Exageración: Describe en qué medida se exagera el tiempo de los acentos.
- Pausas corrientes: Describe el tiempo de la pausa normal.
- Pausas de indecisión: Cantidad de pausas sin significado semántico.
- Ratio de habla: Número de sílabas dichas por segundo.
- Frecuencia de acento: Ratio de palabras acentuadas en la expresión.

[2] "https://es.wikipedia.org/wiki/Síntesis_de_habla"

**Calidad de la voz**

- Respiraciones.
- Brillo de la voz.
- Paso de la voz por la laringe: Que genera un sonido chirriante.
- Volumen: Como resultado de la presión de aire
- Pausas discontinuas: Pausas abruptas o delicadas.
- Cambio discontinuo de la frecuencia.
- Temblor en la dicción.

**Parámetros de articulación**

Precisión en la duración de los fonemas.

## 3.3. Espacio de emociones

En el estudio de PHYSTA se define el concepto de emoción con dos ideas principales. La primera es la idea básica de emoción, originada por Descartes, que dice que los humanos tienen un número concreto de emociones universales, esas emociones se pueden mezclar consiguiendo con ello emociones complejas. La segunda es que la biología motive las emociones, o lo que es lo mismo que las emociones básicas estén motivadas por una necesidad evolutiva.

La clasificación de emociones se puede hacer bajo el punto de vista de que las emociones son discretas o que son continuas. Sin embargo Murray y Arnott (1993)[17] exponen que los dos puntos de vista no tienen porque excluirse el uno al otro ya que los valores discretos pueden ser un punto dentro de los continuos.

Antes de definir los espacios debemos ver a que nos referimos con emoción. A veces el concepto de emocion incluye todos los tipos de emociones, actitudes y conviciones, y otras veces las distingue.

Witcmann(2000) diferencia entre actitud y emociones, para él, las actitudes están menos gobernadas por el sentimiento y más basadas en las convicciones, y una emoción esta basada únicamente en el sentimiento. Por ejemplo la confianza sería una actitud mientras que el miedo sería una emoción. En algunos idiomas las actitudes son declaradas morfológicamente pero no hay ningún idioma que distinga las emociones, o por lo menos que conociera Saeed en 1995.

La mayoría de los investigadores no tienen en cuenta esta diferenciación ya que es muy descriptivo, pero no necesario para categorizar lo que los humanos perciben.[4]

## Categorización discreta de las emociones o EMOCIONES BÁSICAS

Una de las cuestiones actuales más relevantes, al mismo tiempo que más controvertidas, en el estudio de la emoción, es la existencia, o no, de emociones básicas, universales, de las que se derivarían el resto de reacciones afectivas. La asunción de la existencia de tales emociones básicas deriva directamente de los planteamientos de Darwin y significaría que se trata de reacciones afectivas innatas, distintas entre ellas, presentes en todos los seres humanos y que se expresan de forma característica (Tomkins, 1962, 1963; Ekman, 1984; Izard, 1977). La diferencia entre las mismas no podría establecerse en términos de gradación en una determinada dimensión, sino que serían cualitativamente diferentes.

Según Izard,(1997) estas emociones son: Placer, interés, sorpresa, tristeza, ira, asco, miedo y desprecio. Sin embargo, Ekman considera las 6 básicas ira, alegría, asco, tristeza, sorpresa y miedo, a las que añadiría posteriormente el desprecio.

## Categorización continua

Habitualmente se entiende por emoción una experiencia multidimensional con al menos tres sistemas de respuesta: cognitivo/subjetivo, conductual/expresivo y fisiológico/adaptativo.

Desde el advenimiento de la psicología científica, ha habido sucesivos intentos por analizar la emoción en sus componentes principales, que permitieran tanto su clasificación, como la distinción entre las mismas. Quizá la más conocida sea la de teoría tridimensional del sentimiento de Wundt (1896), que defiende que éstos se pueden analizar en función de tres dimensiones: agrado-desagrado, tensión-relajación y excitación-calma.

A partir del planteamiento de Wundt, se han propuesto diferentes dimensiones que caracterizarían las emociones (Schlosberg, 1954; Engen, Levy y Schlosberg, 1958). No obstante, las únicas dimensiones que son aceptadas por prácticamente todos los autores y, que además son ortogonales, son la dimensión agrado-desagrado y la intensidad de la reacción emocional (Zajonc, 1980). Aunque atendiendo únicamente a éstas no puede establecerse una clasificación exhaustiva y excluyente de todas las reacciones afectivas, puesto que emociones como la ira o el odio pueden ser desagradables e intensas y no se trata del mismo tipo de emoción.

Finalmente Oatley (1986) concluye que una emoción podría definirse como una experiencia afectiva en cierta medida agradable o desagradable, que supone una cualidad fenomenológica característica y que compromete tres sistemas de respuesta: cognitivo-subjetivo, conductual-expresivo y fisiológico-adaptativo.

Así, a la hora de emular emociones, generalmente se usan tres dimensiones. Uno de los espacios más usados es el que tiene como ejes fuerza, valencia y actividad (Stibbard (2000) Murray & Arnott (1993)). Un segundo espacio lo componen: nivel de excitación, valencia y la actividad y un tercero :el placer, la excitación y el poder([21]).

También existen sistemas de dos dimensiones que son activo-pasivo(intensidad), y positivo-negativo(agrado-desagrado), así funciona por ejemplo el sistema FEELTRACE.

## 3.4. El espacio continuo de Arousal, Valence, Power

De acuerdo con la teoría de Osgoof, Suci y Tannanbaum(1957) y sucesivos estudios, existen tres principales componentes. [5]

**Arousal** Se refiere a la activación (desde dormido a frenético).

**Valence** Si la intención es positiva o negativa, podríamos interpretar como la evaluación del sentimiento.

**Power** Se relaciona con el poder o sensación de seguridad ante la situación. Nos permite distinguir entre sentimientos similares, como por ejemplo desprecio y miedo.

El plano Power = 0 se representa en la figura 3.4.1

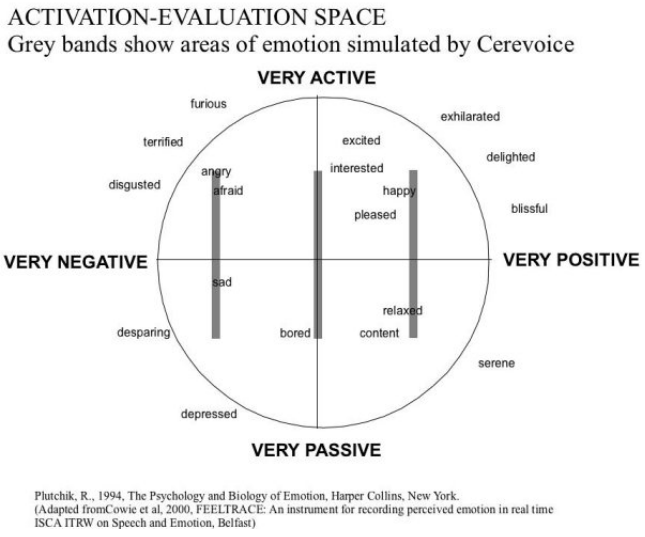



Figura 3.4.1: Plano del espacio tridimensional de emociones definido por 'Power = 0'

A este plano le añadimos otra componente ortogonal (Power). Este eje distinguiría, por ejemplo, miedo y espanto. El miedo tiene un nivel de control de la situación mucho mayor que espanto. Por tanto, en miedo la componente 'power' será menor que en espanto.

Las emociones se representaran con una cantidad entre [-100;100] para cada uno de los 3 ejes. La siguiente tabla muestra las coordenadas de 5 emociones básicas.[27]

| Emoción | Arousal | Valence | Power |
|---|---|---|---|
| Neutro | 0 | 0 | 0 |
| Triste | -8.5 | -42.9 | -55.3 |
| Enfadado | 34.6 | -34.9 | -33.7 |
| Asustado | 31.1 | -27.1 | -79.4 |
| Felíz | 28.9 | 39.8 | 12.5 |

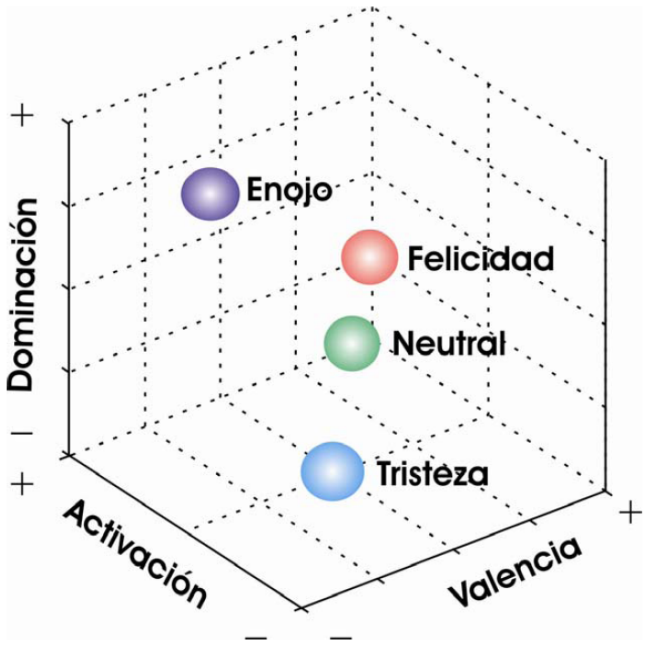



Figura 3.4.2: Ejemplos de emociones en las coordenadas dfinidas

Con este sistema podremos matizar las emociones. Por ejemplo, si a enfadado le aumentamos la coordenada “arousal” la emoción que obtenemos es “ira”.

Este tipo de sistema de categorización emocional nos permite trabajar con infinitas emociones. Esto nos hace mucho mayor el marco expresivo con el que trabajar a hora de la creación de habla emotiva, ya que no nos encontramos encorsetados en unas cuantas emociones..

# Capítulo 4

# Descripción del proyecto

## 4.1. Objetivos de diseño

Nuestro objetivo es dotar a la voz robótica de la característica prosódica. No buscamos una voz natural que se pueda confundir con la humana, buscamos que la voz represente la emoción y esto acompañe al mensaje, obteniendo todas las ventajas explicada anteriormente.

Dotaremos al sistema de intención expresiva, es decir, si la frase es enunciativa, exclamativa o interrogativa, y expresión emocional.

El objetivo, en cuanto a la expresión emocional, es ser capaz de expresar cualquier emoción dentro del espacio de emociones definido. Esto implica tener un espacio continuo en el cual los parámetros de voz utilizados se correspondan de manera continua con los parámetros emocionales.

Existen varios metodos para cumplir los objetivos expuestos. A continuación vamos a concretar las tecnologías usadas para la consecución de nuestros objetivos en los siguientes apartados:

- Elección de parámetros para lingüísticos.
- Elección del sistema de síntesis

## 4.2. Elección de los parámetros paralingüísticos

Para simplicidad del sistema hemos cogido solo algunos de estos parámetros, la selección se ha hecho comparando estudios [13, 19, 22, 24, 26] y observando cuales son los efectos que más afectan al cambio en las emociones.

**Parámetros de frecuencia**

- Media de frecuencia: Es la base de la entonación, diferencia emociones directamente.
- Rango de frecuencia: El movimiento más o menos amplio es muy representativo en varías emociones.
- Final de la expresión: Si la frase acaba subiendo o bajando. Nuestro diseño hará que pueda subir o bajar más o menos rápido de forma que haya diferentes pendientes para diferentes puntos en el espacio emocional.
- Línea de referencia: Línea que genera la entonación (enunciativa, interrogativa o exclamativa).

**Parámetros de tiempo**

- Pausas: Uniremos las pausas corrientes y las de indecisión en un solo tipo de pausa, pero en las emociones más "inseguras" las pausas serán más largas
- Ratio de habla: Este parámetro es un parámetro controvertido, ya que, al cambiarlo, la inteligibilidad del sistema varía notablemente, pero es de los más importantes a la hora caracterizar una emoción.
- Variación del ratio de habla. Varía el ratio del habla para distintas sílabas, lo que permite el habla desigual, que da la sensación de inseguridad.

**Calidad de la voz**

De la calidad de la voz mantenemos pocos parámetros, ya que, son más complicados de implementar y generar una norma.

- Volumen: Como resultado de la presión de aire
- Cambio discontinuo de la frecuencia que generamos a partir de la frecuencia de acento o probabilidad de tener una letra acentuada

**Parámetros de articulación**

Precisión en la duración de los fonemas que se puede emular con la variación del ratio de habla.

## 4.3. Elección del sistema de síntesis

Las necesidades del diseño son:

- Modificar frecuencia y duración.
- Alta inteligibilidad.
- Poca peso computacional, ya que necesitamos que las frases se creen casi a tiempo real.
- Poco tamaño de base de datos.
- Software libre para poder usar librerías hechas ya y, además, poder contribuir con nuestro proyecto.

Teniendo en cuenta estas características comentadas en el apartado 3.1, decidimos según requerimientos técnicos usar síntesis concatenativa por difonos con Mbrola.

Mbrola es un sintetizador basado en la concatenación de dífonos, que usa la tecnología de Psola.

No es un TtS ya que no acepta la entrada de texto directo si no de fonemas. La entrada es una lista de fonemas junto con su información prosódica (duración de los fonemas y una descripción por tramos de la frecuencia), a la salida, produce muestras de habla de 16bits(lineal), con la frecuencia de muestreo de la base de datos usada.

Este es un sintetizador libre solo para aplicaciones no comerciales ni militares.

> "MBROLA success is much more the result of a dynamic and collaborative community than a strong technical improvement."
> Prof. Thierry Dutoit
> TCTS Lab - FPMs

El proyecto Mbrola surge como una iniciativa del TCTS Lab de la Faculté Polytechnique de Mons (Belgica), su objetivo es obtener un conjunto de sintetizadores de lenguaje en el mayor número de lenguas posible y suministrarlas de manera libre para aplicaciones no comerciales. El fin último es impulsar la investigación académica en síntesis de voz, en particular de generar prosodia.

**¿Por qué una base de difonos propia?** Esta es otra de las partes interesantes del proyecto, además de generar nuestro propio sistema, creamos una base de difonos en español con timbre de femenino al proyecto. Esta base de datos podrá ser usada por Mbrola y por otros sintetizadores de voz más modernos (Festival) en cualquier lugar del mundo.

Con esto ponemos un granito de arena en el desarrollo técnico de esta rama. Mbrola cuenta ya con bases de datos en 36 idiomas, desde el griego clásico hasta el mahorí, diferenciando el masculino y el femenino. Nuestra base de datos será la primera en Español femenino.

## Funcionamiento general de Mbrola.

Como citamos brevemente en la sección 4.1, Mbrola es un software de síntesis de voz concatenativa.

Podemos descomponer el funcionamiento de Mbrola en 4 pasos[1]:

- El archivo . pho
- Selección de la base de datos.
- Procesado PSOLA
- Concatenación con alisado de bordes.

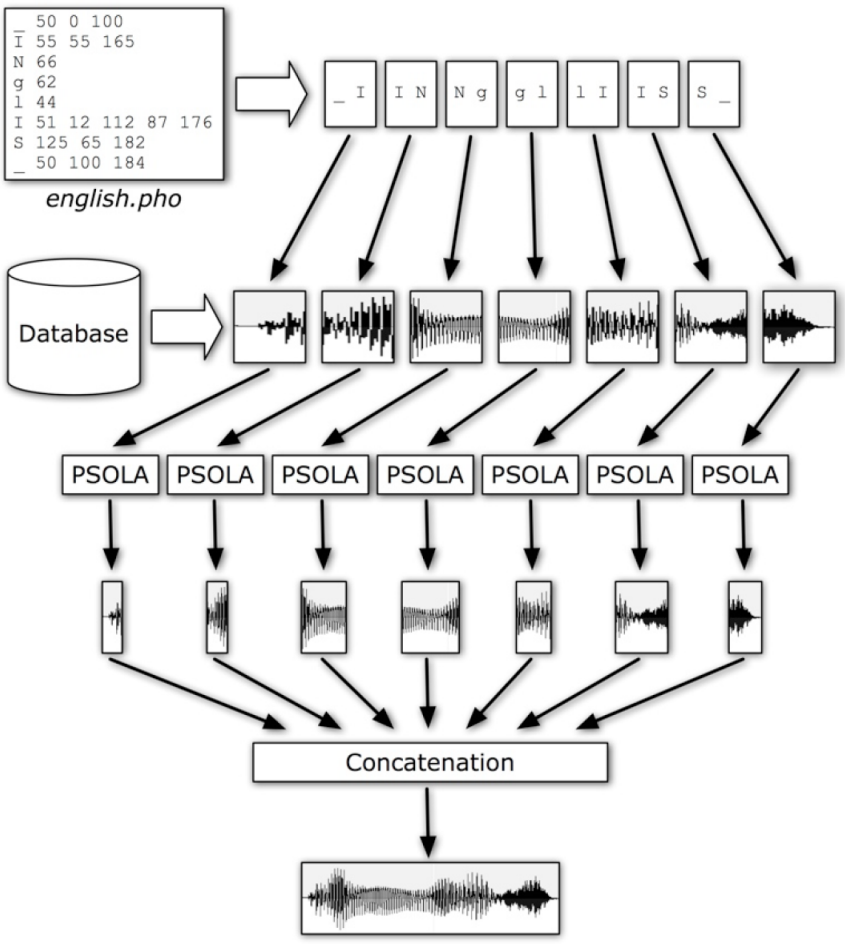



Figura 4.3.1: Descripción de la formación de la frase con Mbrola.([1])

**Ficheros de entrada a Mbrola (.pho)**

Son el tipo de archivo del que coge Mbrola la información para crear la frase. Este contiene la información fonética y de entonación de la frase.

Se organiza en filas; cada una de ellas contendrá los fonemas a representar, después el tiempo en ms que dura en total, y después desarrolla como va a ser la forma de la frecuencia. Por ejemplo, si queremos que dure 100ms y que al principio esté a 440Hz luego a la mitad de la duración a 420 y al final vuelva a 440 y es una 'a'. Esta línea del archivo sería así:

a 100 0 440 50 420 100 440.

Las uniones entre un punto y otro se hacen linealmente y así podemos modelar la frecuencia en cada uno de los fonemas que forman la frase.

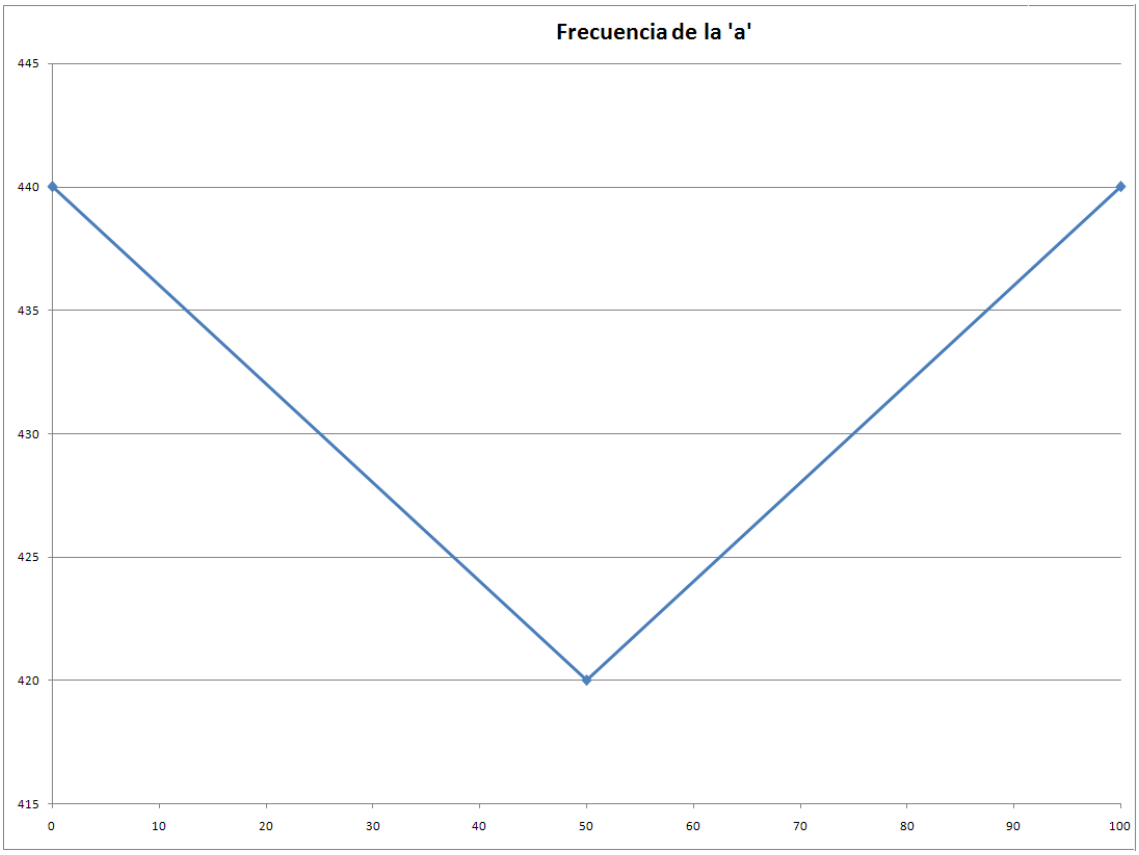



Figura 4.3.2: Frecuencia(Hz); Tiempo (ms). Formación del tono a partir una línea del fichero "fonético".

En las siguientes líneas tenemos representados

_ 23 0 49

a 163 22 41 59 40 96 37

o 163 33 45 70 45 100 47

r 94 11 56 75 45

a 149 25 61 65 68 100 84

_ 25 31 80

El anexo 2 podemos encontrar un fichero .pho completo.

**Selección de la base de datos.**

Una vez separados los dífonos se procede a su búsqueda en nuestra base de datos de sonidos, según se muestra en la siguiente figura.

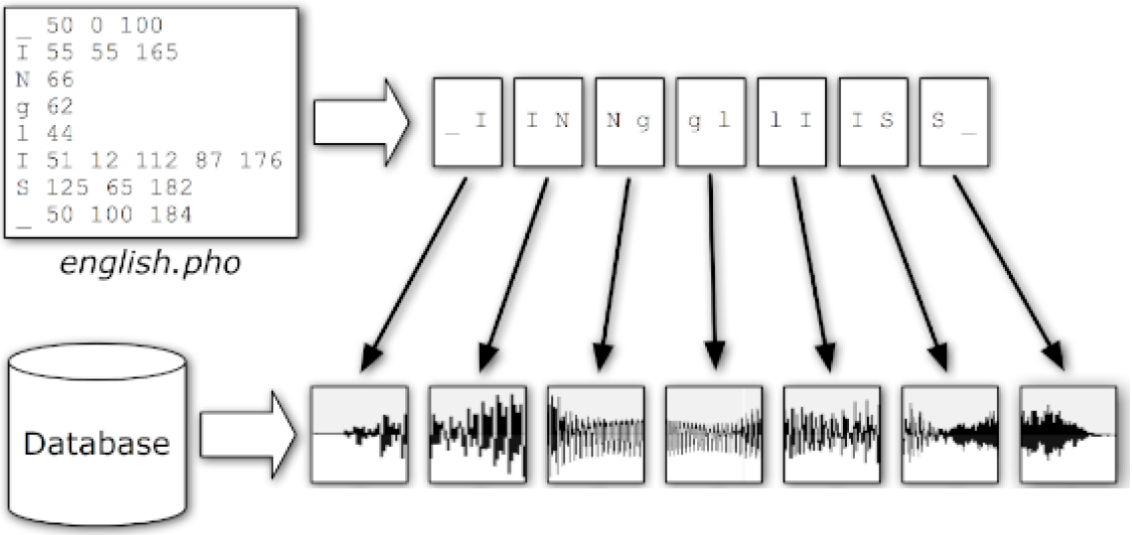



Figura 4.3.3: Recuperación del sonido de la base de datos.([1])

**El procesado Psola.**

(Pitch Synchronous Overlap and Add) [7]es una técnica de procesado de señal usado para el procesado de voz, más concretamente, para la síntesis del habla. Puede ser utilizado para modificar la frecuencia y la duración de la señal de habla.

PSOLA funciona dividiendo la onda en pequeños segmentos superpuestos. Para cambiar la frecuencia de la señal los segmentos son alejados (más graves) o acercados(más agudo)

para cambiar la frecuencia. Para cambiar la duración de la señal los segmentos se repiten varias veces(alargar) o son eliminados(acortar).

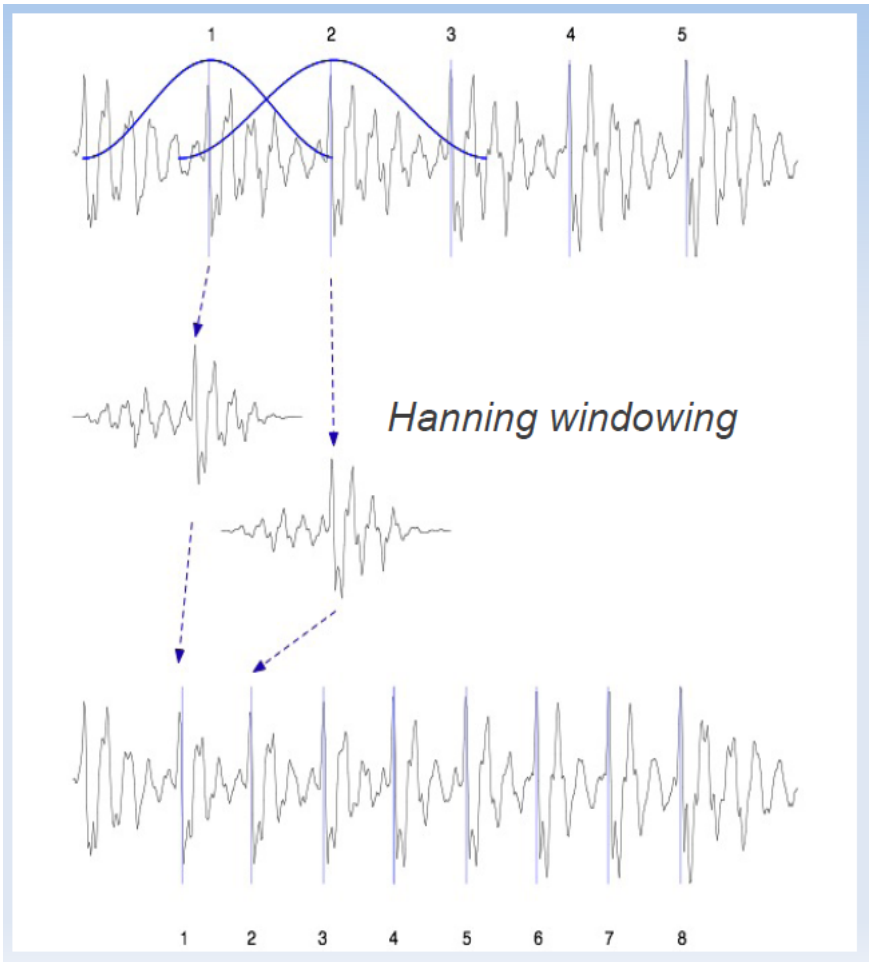



Figura 4.3.4: Funcionamiento de Psola.[7]

**Suavizado de bordes.**

El proceso acaba suavizando las uniones entre los audios de los distintos difonos. De esta manera disminuimos los ruidos( clicks ) causados por 3 tipos de discontinuidad : De fase, de tono y de espectro.

# Capítulo 5

# Creación de base de datos de difonos en español con voz femenina.

La base de datos de MBROLA es una recopilación de todos los difonos usados en el idioma, es decir, de todos los difonos los que nos permitan crear cualquier frase en español.

El difono es el sonido que va desde la región estable de un fono[1] a la región estable del siguiente.Los difonos contienen la información de la articulación entre fonos y disminuyen los saltos de la unión ya que esta es hace por la zona estable.

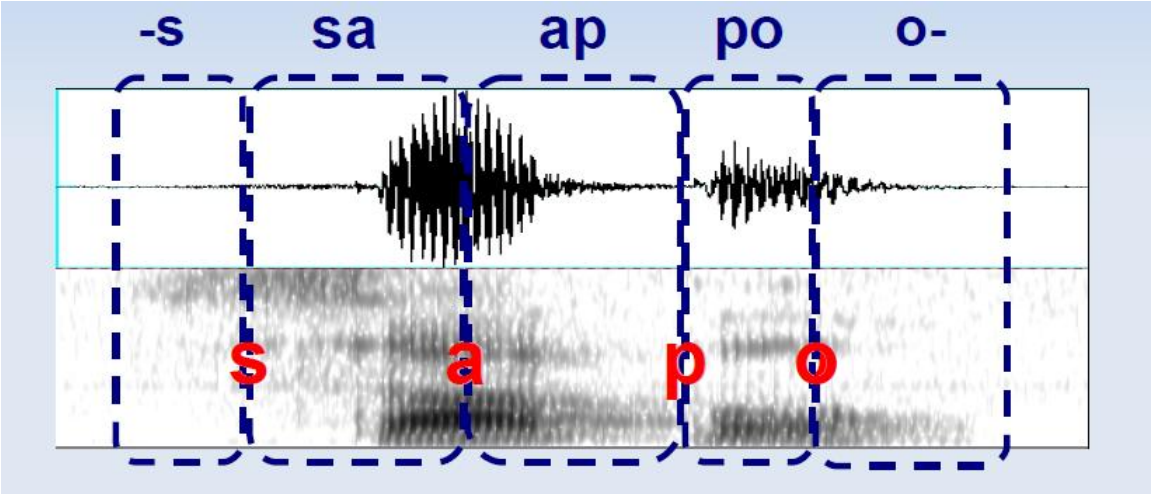



Figura 5.0.1: Selección de difono en la señal de audio[7]

Para crear la base de datos vamos a seguir los siguientes pasos:

- Proponer una tabla de difonos posibles en el idioma.
- Hacer una colección de los difonos usados buscando palabras que los contengan.

[1]Fono es cada uno de los segmentos de características acústicas particulares en que podemos dividir la secuencia sonora. Se representan por medio de fonemas. http://es.wikipedia.org/wiki/Fono

- Elaborar frases con esas palabras.
- Grabar las frases.
- Cortar los dífonos y señalar las zonas estables.
- Generar la base de datos del tipo Mbrola.

**Elaboración del texto a grabar.**

La elaboración del corpus incluye los 3 primeros pasos de la creación de la base de datos.

Creamos frases que contengan los dífonos del castellano usado en España (más concretamente en Madrid). La localización es importante porque la fonética varía según la región.

En nuestra base de datos asumimos que 'b' y '*v*' tienen el mismo sonido, así como '*y*' y 'll', por tanto no diferencimos entre 'b' y 'v' ni entre 'll' e 'y'. Tampoco vamos a tener en cuenta los alófonos[2].

Para buscar todos los dífonos usados en español concatenamos todos los fonemas de dos en dos y buscamos palabras con cada uno de ellos. Los dífonos de los que no encontramos palabra se consideran como no utilizados. En la siguiente tabla se muestran las combinaciones de fonemas usados para nuestra base de datos.

En la filas se representa el primer fonema y en la columna el segundo. Las letras representan su sonido en castellano, no usamos el sistema SAMPA porque es más complicado a la hora de implementar.

La leyenda es la siguiente:

**'_'** Principio o final de palabra

**'b'** Representa el sonido de b y v.

**'g'** Representa el sonido de la g seguida de la u. Por ejemplo gue o gui.

**'j'** Representa sonido de la j como por ejemplo en jamón.

**'rr'** Representa a la r rodada.

**'ll'** Representa tanto la ll como la y.

El resto de letras representan su sonido en castellano.

[2]Los alófonos son sonidos que son reconocidos en determinado lenguaje como el mismo fonema. Por ejemplo, en español, la 'd' se reconoce con distinto sonido en distintas ocasiones. Las dos 'd's de "dado" son sonidos distintos.

@ Indica que esa concatenación de fonemas no se encuentra en ninguna palabra para el castellano en España.

| | _ | a | b | k | d | e | f | g | i | j | l | m | n | ñ | o | p | r | rr | s | t | u | ch | ll | z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **_** | @ | _a | _b | _k | _d | _e | _f | _g | _i | _j | _l | _m | _n | _ñ | _o | _p | _r | _rr | _s | _t | _u | _ch | _ll | _z |
| **a** | a_ | aa | ab | ak | ad | ae | af | ag | ai | aj | al | am | an | añ | ao | ap | ar | arr | as | at | au | ach | all | az |
| **b** | b_ | ba | @ | @ | bd | be | @ | @ | bi | bj | bl | @ | @ | @ | bo | @ | br | @ | bs | bt | bu | @ | @ | @ |
| **k** | k_ | ka | @ | @ | @ | ke | @ | @ | ki | @ | kl | @ | @ | @ | ko | @ | kr | @ | ks | kt | ku | @ | @ | kz |
| **d** | d_ | da | db | dk | @ | de | @ | @ | di | @ | dl | dm | @ | @ | do | @ | dr | @ | @ | dt | du | @ | @ | @ |
| **e** | e_ | ea | eb | ek | ed | ee | ef | eg | ei | ej | el | em | en | eñ | eo | ep | er | err | es | et | eu | ech | ell | ez |
| **f** | @ | fa | @ | @ | @ | fe | @ | @ | fi | @ | fl | @ | @ | @ | fo | @ | fr | @ | @ | ft | fu | @ | @ | @ |
| **g** | @ | ga | @ | @ | @ | ge | @ | @ | gi | @ | gl | @ | gn | @ | go | @ | gr | @ | @ | gt | gu | @ | @ | @ |
| **i** | i_ | ia | ib | ik | id | ie | if | ig | ii | ij | il | im | in | iñ | io | ip | ir | irr | is | it | iu | ich | ill | iz |
| **j** | @ | ja | @ | @ | @ | je | @ | @ | ji | @ | @ | @ | @ | @ | jo | @ | @ | @ | @ | jt | ju | @ | @ | @ |
| **l** | l_ | la | lb | lk | ld | le | lf | lg | li | lj | @ | lm | @ | @ | lo | lp | @ | @ | ls | lt | lu | lch | @ | lz |
| **m** | m_ | ma | mb | @ | @ | me | @ | @ | mi | @ | @ | @ | mn | @ | mo | mp | @ | @ | ms | mt | mu | @ | @ | @ |
| **n** | n_ | na | @ | nk | nd | ne | nf | ng | ni | nj | @ | nm | @ | @ | no | @ | @ | nrr | ns | nt | nu | nch | nll | nz |
| **ñ** | @ | ña | @ | @ | @ | ñe | @ | @ | ñi | @ | @ | @ | @ | @ | ño | @ | @ | @ | ñs | ñt | ñu | @ | @ | @ |
| **o** | o_ | oa | ob | ok | od | oe | of | og | oi | oj | ol | om | on | oñ | oo | op | or | orr | os | ot | ou | och | oll | oz |
| **p** | p_ | pa | pb | pk | @ | pe | @ | @ | pi | @ | pl | pm | @ | @ | po | @ | pr | @ | ps | pt | pu | @ | @ | pz |
| **r** | r_ | ra | rb | rk | rd | re | rf | rg | ri | rj | rl | rm | rn | @ | ro | rp | @ | @ | rs | rt | ru | rch | @ | rz |
| **rr** | @ | rra | @ | @ | @ | rre | @ | @ | rri | @ | @ | @ | @ | @ | rro | @ | @ | @ | rrs | rrt | rru | @ | @ | @ |
| **s** | s_ | sa | sb | sk | @ | se | sf | sg | si | @ | sl | sm | sn | @ | so | sp | @ | @ | @ | st | su | sch | sll | sz |
| **t** | t_ | ta | @ | ? | @ | te | @ | @ | ti | @ | tl | @ | tn | @ | to | @ | tr | @ | ts | @ | tu | @ | tll | @ |
| **u** | u_ | ua | ub | uk | ud | ue | uf | ug | ui | uj | ul | um | un | uñ | uo | up | ur | urr | us | ut | uu | uch | ull | uz |
| **ch** | ch_ | cha | @ | @ | @ | che | @ | @ | chi | @ | @ | @ | @ | @ | cho | @ | @ | @ | @ | @ | chu | @ | @ | @ |
| **ll** | @ | lla | @ | @ | @ | lle | @ | @ | lli | @ | @ | @ | @ | @ | llo | llp | @ | @ | lls | @ | llu | @ | @ | llz |
| **z** | @ | za | @ | zk | @ | ze | @ | @ | zi | @ | @ | @ | @ | @ | zo | @ | @ | @ | zs | @ | zu | @ | @ | @ |

Figura 5.0.2: Dífonos existentes en palabras en castellano.

Además de estos dífonos es necesario grabar la unión entre palabras, ya que al hablar, muchas veces unimos una palabra con la siguiente.

Para ello, de los dífonos que no hemos considerado (los que se representan como @ en la figura 5.0.2) vamos a incorporar los que puedan formarse al unir dos palabras. Tomamos como fonemas iniciales del dífono 'j','l','m','p','r','z' (final de la primera palabra) y los concatenamos con todos los fonemas finales (principio de la segunda palabra) que falten por unir. Con estos fonemas se construyen uniones de nombre y adjetivo. Hemos usado nombre y adjetivo para mantener un sentido gramatical y que de esta manera sea más fácil de grabar. Por ejemplo. "reloj bonito" para el dífono 'jb', "reloj nuevo" 'jn'. De esta manera añadimos los dífonos que nos faltan.

El conjunto de palabras debe tener todos los dífonos en el menor número de palabras, de forma que minimicemos el tiempo, tanto en grabar como en recortar. Por ejemplo, de 'causa' podemos aprovechar _-k k-a a-u u-s s-a a-_ . No debemos entonces grabar 'casa' ya que tenemos esos dífonos en la otra palabra. Aun así, es recomendable tener, por lo menos, dos ejemplos de cada dífono en la grabación por si alguno estuviera dañado.

Es recomendable no grabar los difonos objetivo al principio o al final de la frase, ya que suenan de manera distinta al empezar y terminar la locución.

Las frases concretas y las palabras que contienen los difonos se encuentran en el Anexo 1.

## Grabación y corte de las frases.

En este apartado vamos a explicar cómo se realiza la grabación del conjunto de frases diseñado.

Para la grabación tenemos que elegir un buen locutor. En nuestro caso necesitamos que sea una voz femenina. La locutora, por tanto, tendrá que vocalizar correctamente (sin llegar a sonar antinatural) y ser capaz de mantener un tono constante durante toda la grabación y hablar de la forma más neutra que pueda, como si de una enumeración se tratara.

Debemos grabar en un lugar con buena acústica, poca reverberación y poco ruido, es decir, acercarse lo más posible a las condiciones de un estudio de grabación.

Para obtener una señal de audio de buena calidad utilizamos un micrófono con alimentación externa PHANTOM y con cápsula de condensador. El micrófono va conectado directamente a la tarjeta de sonido del equipo ASUS Xonar D-KARA funcionando con el driver VIA HD versión 7.3.00.30.

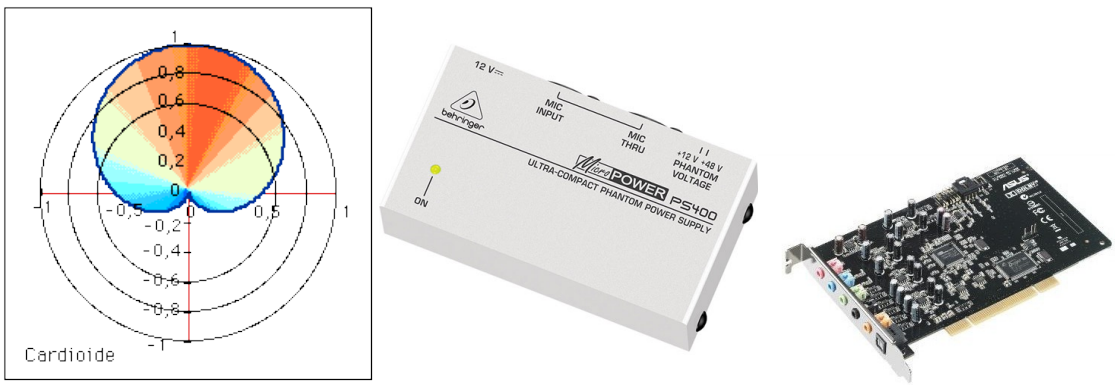



Figura 5.0.3: Equipo de grabación: Diagrama polar del micrófono, alimentador del PHANTOM y tarjeta de sonido.

El Software usado para grabar y cortar los audios es el programa gratuito Audacity[3]. Configuramos la línea de grabación a 16Khz, 32-bit sin signo y mono, para compatibilizarlo con Mbrola.

Grabamos las frases en una de las pistas y recortamos cada una de las palabras que vamos a utilizar. Una vez llevadas a una pista distinta silenciamos el resto y las palabras separadas

[3] audacity.sourceforge.net/

se exportan. La palabras las guardamos en un archivo que nombraremos con la palabra que contiene ( “perro.wav”). Los dífonos que salen de unir palabras se guardan con las dos palabras juntas.

Hay que tener cuidado con la duración de los cortes de las palabras ya que el programa que usaremos en el paso siguiente tiene dos restricciones:

- Un tiempo máximo de duración de archivo
- Necesitas dejar un margen desde el principio del audio hasta el inicio del dífono objetivo, y equivalentemente al final.

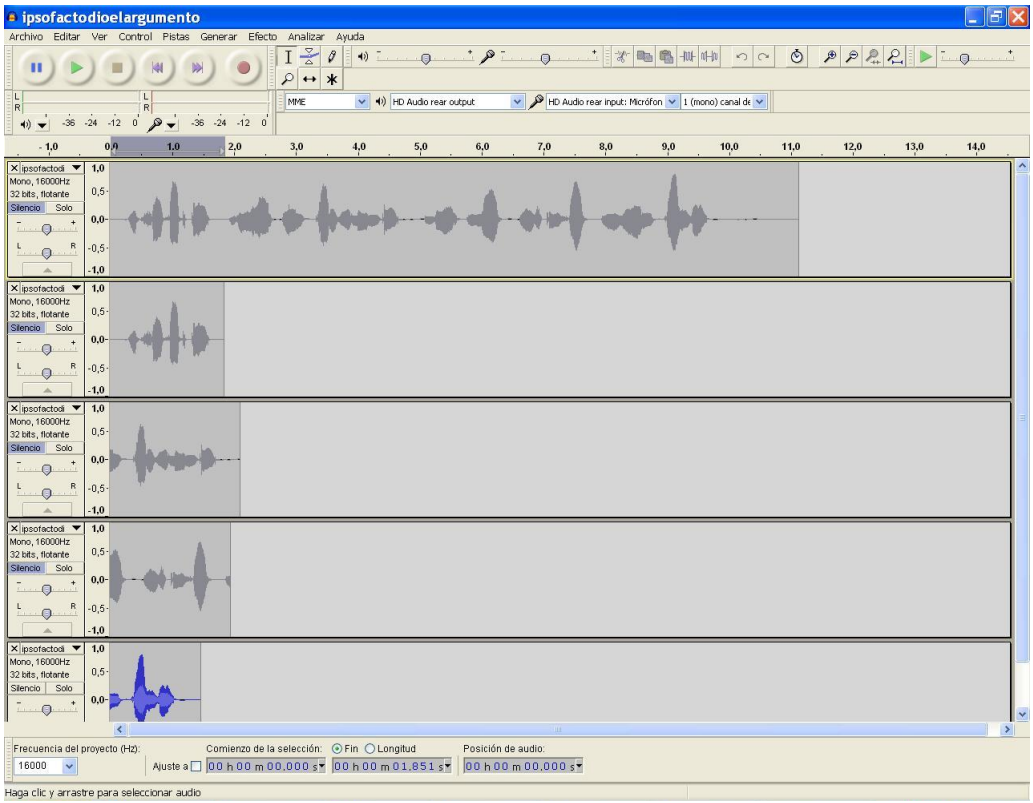



Figura 5.0.4: Pistas en Audacity

## Creando la base de datos

Una vez separadas las palabras en audios distintos necesitamos organizar la información de manera que la pueda procesar Mbrola. Necesitemos que esta tenga el siguiente formato:

```
!16000
_       a       apresuradamente.wav       <UTTERANCE> 15423 17058 17936
_       b       burbuja.wav       <UTTERANCE> 5566 7120 8258
_       d       dueto.wav       <UTTERANCE> 6530 8134 9245
_       e       enun.wav       <UTTERANCE> 7814 9063 11625
_       f       futbol.wav       <UTTERANCE> 2384 2574 4924
_       g       ganan.wav       <UTTERANCE> 3673 5103 5823
_       H       chema.wav       <UTTERANCE> 8538 9158 10218
_       i       izquierda.wav       <UTTERANCE> 3230 4500 5490
_       j       juan.wav       <UTTERANCE> 9298 10978 11908
_       k       canoa.wav       <UTTERANCE> 1776 2416 2736
_       l       lee.wav       <UTTERANCE> 7410 7987 9117
_       m       mancha.wav       <UTTERANCE> 3481 3971 5281
_       n       naftalina.wav       <UTTERANCE> 2949 3559 4919
_       N       nonerias.wav       <UTTERANCE> 5786 7036 8336
....
```

En la primera y segunda columna tenemos la representación de los fonos que forman el dífono. En la tercera, el archivo donde encontramos el sonido. Le sigue una columna de comentarios. Los siguientes tres números indican el frame en el que realizaremos los cortes: El primero es el frame donde se encuentra la parte estable del primer fono que forma el dífono, el tercero es la parte estable del segundo fono y el número del medio es el frame en el que el dífono cambia de un fono a otro.

Podríamos hacer el archivo a mano, pero *Arthur Dirksen* desarrolló una aplicación, *Diphone Studio software*[4], para facilitar el trabajo de encontrar los frames. Esta aplicación permite hacer la separación indicando los puntos de manera visual y guardándolos directamente ordenados como Mbrola requere.

El programa necesita saber en qué archivo encontrará cada dífono, para esto les damos los datos ordenados de esta manera:

[4]http://www.fluency.nl/dstudio/dstudio.htm

l p culpa.wav <UTTERANCE>

l R arbolr.wav <UTTERANCE>

l s solsticio.wav <UTTERANCE>

l t altavoz.wav <UTTERANCE>

l u lugar.wav <UTTERANCE>

l y arboly.wav <UTTERANCE>

l z calzado.wav <UTTERANCE>

m _ referendum.wav <UTTERANCE>

m a chema.wav <UTTERANCE>

m b bambu.wav <UTTERANCE>

m d referendumd.wav <UTTERANCE>

Las primeras dos columnas serán los dos fonemas, en orden, que forman el difono. La tercera es el archivo de audio que contiene el difono objetivo. La cuarta es para aclaraciones o notas.

Una vez que tenemos todos los difonos ordenados guardamos el archivo con extensión ".lst" (en nuestro caso vamos a llamarle "base.lst").

Para iniciar el proceso, vamos a "File/New" y seleccionamos nuestro archivo. Después vamos a "File/Open" para crear un archivo con extensión ".dat" que deberá colocarse en el mismo directorio.

La primera línea de esa base de datos debe incluir la frecuencia a la tenemos grabadas las frases (Por ejemplo !22050 quiere decir que esta grabado a 22050 muestras por segundo) por defecto la frecuencia es 16000.

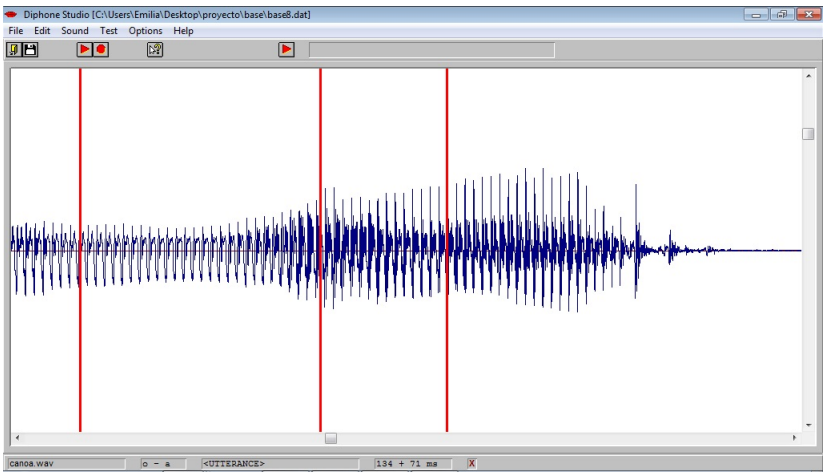



Figura 5.0.5: Recorte de difonos con DSTUDIO

Ahora, con precisión, tenemos que marcar las zonas estables y el cambio de fono.

La pantalla del programa mostrará en este orden (como vemos en la figura anterior):

- La forma de onda de archivo sobre el que estamos trabajando.
- El nombre del archivo del que la contiene.
- El difono objetivo
- Los comentarios para ese difono
- La distancia entre los tres separadores.

Cuando guardemos, el programa almacernará el número de frames en la base de datos. Habrá que realizar esto con cada difono.

El software nos permite probar uniendo difonos entre ellos para ver si es inteligible la unión de distintos difonos. Hay que tener en cuenta que el programa no realiza el suavizado de las uniones, por esto, habrá discontinuidades.

Ahora tenemos una estructura muy similar a la que necesita el software con el que generaremos la base de datos.

## Hacia la base de datos de con la herramienta de MBrola.

Para obtener la DB(Data Base) primero tenemos que reorganizar las columnas de la siguiente manera:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| p | | e | | perro.wav | 1596 | 1600 | | 1654 | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| perro.wav | | p | e | | 1596 | 1654 | | 1600 | |

Renombramos el archivo “base.dat” y lo llamamos “base.seg” y generamos un directorio con los archivos de audio que se llamará “AUDIO”.

Vamos a necesitar resintetizar la voz para generar la base de datos. Para eso los estudiamos y resintetizamos con el software de Mbrola como se indica en la siguiente figura:

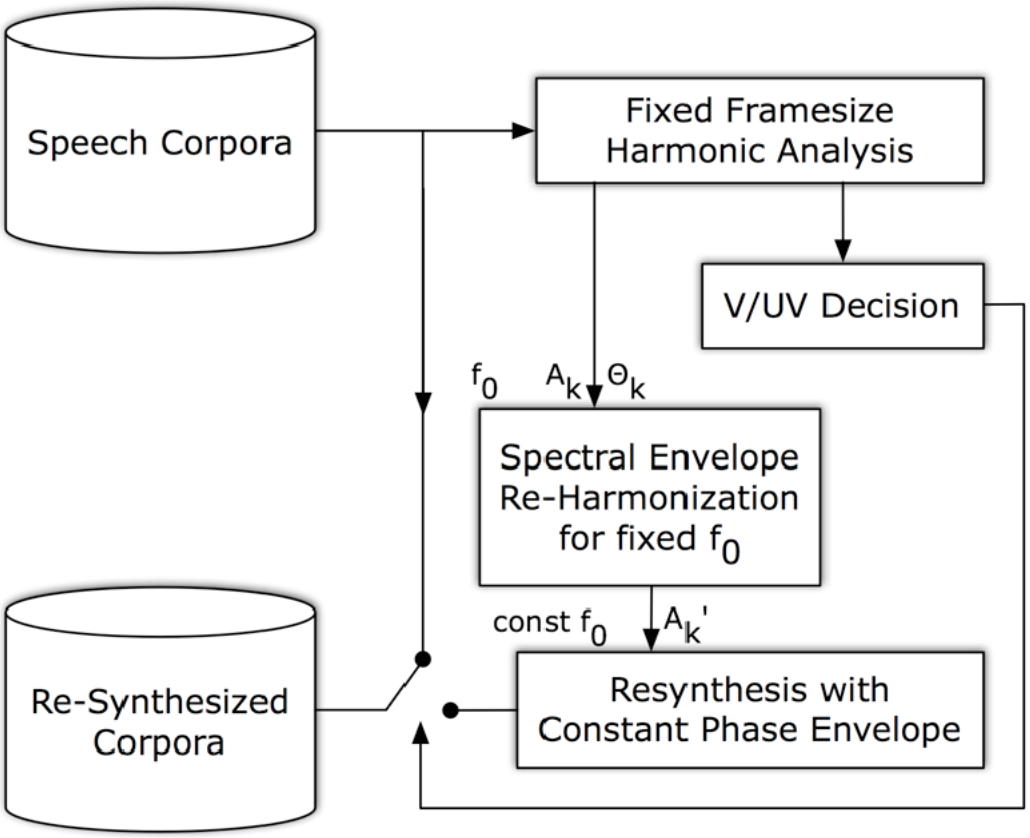



Figura 5.0.6: Estudio y resíntentesis de cada uno de los audios[1].

Utilizamos para esto los siguientes comandos en un terminal linux que contenga los ejecutables que proporciona Mbrola para crear sus bases de datos:

Dentro de esa carpeta ejecutamos “generate_mbrola base.seg” (siendo base.seg el nombre de nuestra base de datos) y esto genera tres archivos:

- “base.f0”, que contendrá datos para el análisis de la frecuencia medio.
- “base.mbe”, que contendrá datos para un análisis de armónicos y ruido.
- “base.syn” que contendrá datos para la resintentización al formato Mbrola.

Antes de seguir debemos modificar el archivo “base.mbe”. En nuestro caso modificamos los valores de estas variables y los sustituimos por “FrameShift” a 432 y “FrameLength” a 144.

Se realizarán ahora esos tres análisis para cada archivo de audio del directorio “AUDIO”. Esto generará un archivo con extensión “.s16” por cada audio como resultado.

Ejecutamos “generate_make.pl” indicando también el nombre de la base, la carpeta con los audios y la carpeta donde poner los resultados (en nuestro ejemplo:“generate_make.pl base AUDIO/ res/ res/”). Generaremos, con esto, en el directorio “res” (donde se crean los

resultados) tres ficheros por cada grabación (Un “.fo”, un “.mbe” y un “.syn”). Y, además, nos genera el archivo “base.mak”.

En general estariamos preparados para obtener nuestra base de datos, pero, en nuestro caso, tenemos que hacer un cambio ya que al ser una voz femenina tenemos que resintetizarla de manera especial. Para eso abrimos el archivo base.mak y sustituimos “resynth” por “resynth_female”.

Ejecutamos el make “make -f base.mak” y así obtendremos entonces nuestra propia base de datos en español con timbre femenino.

Una vez que la tenemos podemos mandarla a Mbrola para compartir nuestro trabajo. Llamaremos a nuestra base durante el proyecto “*Spanishfemale*”.

# Capítulo 6

# Generación de voz emotiva a partir del software de Mbrola.

En este capítulo contaremos cómo resolvemos el problema de la voz emotiva con Mbrola.

Como ya hemos explicado en 4.3 .Mbrola coge la información de un archivo con extensión .pho. Este archivo contiene la información de:

- Los fonemas que vamos a usar.
- La frecuencia de esos fonemas en unos momentos determinados.

Pues bien, esta parte del trabajo, consiste en convertir un punto emocional (una emoción) y un texto en un archivo extensión ".pho". Después Mbrola convertirá ese archivo en un audio y finalmente el reproductor del sistema lo reproducirá.

En la siguiente figura se expone su funcionamiento.

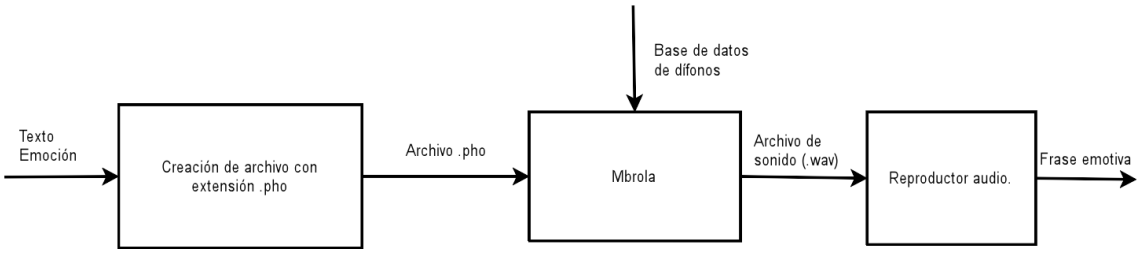



Figura 6.0.1: Diagrama general del programa.

El programa se ejecuta con la siguiente sentencia.

/EmoSoft <texto> <0/1 la máscara neutra> < emoción> <nombre audio de salida>

A continuación explicaremos cada uno de los parámetros

1. Ruta del fichero que contiene el texto a reproducir

2. Activación de "máscara neutra": si está activada (valor=1), se genera un audio sin emoción.

3. La emoción viene representada por los tres parámetros siguientes. Son tres números reales entre menos uno y uno, ambos incluidos [-1;1] y representa a la emoción de la con las tres coordenadas siguientes (Arousal, Valence, Power)

4. Por último ponemos el nombre del archivo de salida, que debe incluir la extensión .wav para luego poder ser reproducido sin problemas.

## 6.1. Generación del fichero de entrada a Mbrola (.pho)

La parte del programa que crea el fichero de entrada a Mbrola funciona de esta manera:

El programa recibe un texto y una emoción, con la emoción se calculan los parámetros prosódicos característicos y obtenemos la configuración ( que es la información de lo paramentros prosódicos). Del texto sacaremos la información sobre cada fono, si es vocal, si esta acentuado, si es un espacio. Uniendo esto con lo parámetros prosódicos de duración de fonos, sacamos la duración de cada fonema y la duración total que tendrá la frase. Al encontrar un signo de exclamación o de interrogación cambiamos una variable que almacena el modo de la comunicación [1],representando, cada uno, con un número entero.

Una vez identificado el modo de la comunicación abrimos una curva modal[2] que se adecue a él. Con este fin tenemos guardado varios archivos con ejemplos de curvas de cada modo. Se tomará aleatoriamente una de las curvas del modo adecuado y la modificamos para que tenga la frecuencia media y la duración requeridas para nuestro mensaje.

Llegados a este punto tenemos una curva modal adaptada a la frecuencia y a la duración de nuestro mensaje y por otra parte, al principio, dimos una duración a cada fono.Con el momento en el que se poduce cada fono y la curva sacamos la frecuencia principal de cada fono. Y llegado a este punto solo nos falta modificar un poco la envolvente de la frase con el resto de paramentros prosódicos calculados.

Este código está estructurado en las funciones que enumeramos a continuación.

1. Cálculo de los parámentros expresivos a partir del punto en el espacio de emociones.

2. Conversión del texto en fonemas y extracción de la información prosódica del texto.

3. Elección de la curva base. (Enunciativo, interrogativo, exclamativo).

[1](enunciativo, exclamativo o interrogativo)

[2]Es la curva que representa la forma de la envolvente de la frecuencia respecto al tiempo.

4. Adaptación del la curva base a nuestra frase objetivo ( En frecuencia y tiempo)

5. Creación del fichero .pho con todos los datos obtenidos anteriormente.

La figura resume el funcionamiento y explica gráficamente el flujo de la información de nuestro programa.

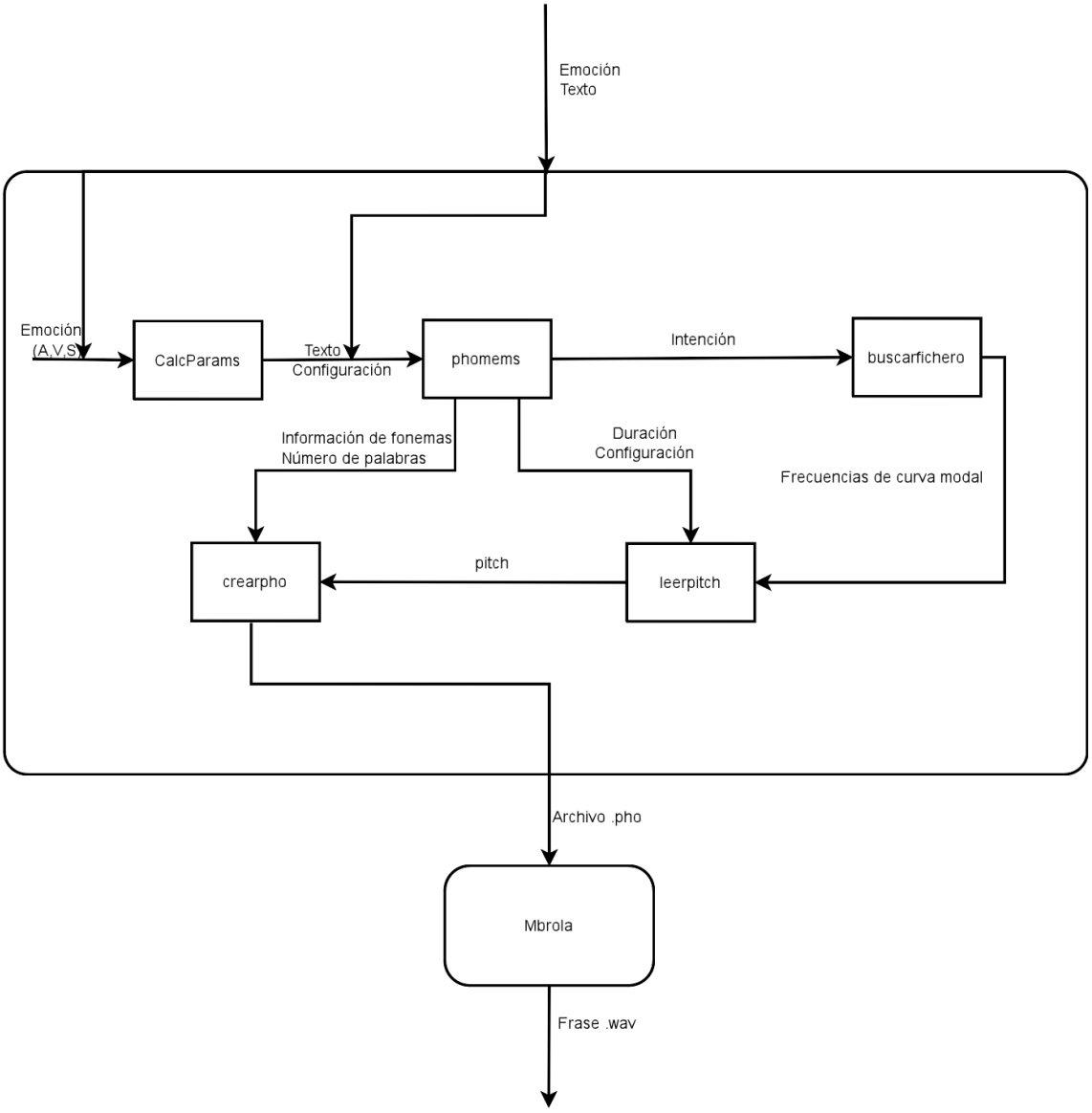



Figura 6.1.1: Diagrama del programa

## Conversión de las coordenada de la emoción a parámetros expresivos (Función CalcParams)

Recoge los datos de la emoción y los convierte en los parámetros prosódicos que usaremos para generar la emoción.

Para esto vamos a basarnos en el artículo de Schroder [27] con ligeras modificaciones en la elección de los parámetros prosódicos. Este artículo estudia las relaciones líneales entre los parámetros expresivos y de las coordenadas de la emoción.

En nuestro proyecto vamos a usar los siguientes parámetros prosódicos.

- Volumen: Es la potencia a la que se reproducirá la frase. Aumenta con "Arousal"
- Duración media(ms): Es la duración media de cada fonema. Aumenta con "Arousal" y disminuye con "Status".
- Variación de la duración(ms): Es lo que varía de un fonema a otro la duración. Disminuye con "Valence" y disminuye con "Status".
- Frecuencia media (Hz): Es la frecuencia media sobre la que se montará la curva.
- Variación de la frecuencia(Hz): Aumenta con "Arousal" y disminuye con "Status".
- Pendiente final(Hz/ms): Es el final de la frase si acaba acentuado o no. Aumenta con "Arousal" y con "Valence". En nuestro programa esto va del -3 al 3 y crecerá, o decrecerá, la última palabra de la frase con más o menos pendiente.
- Duración de las pausas (ms): Aumenta la duración de las pausas respecto a la de los fonemas. Disminuye con "Valence" y con "Arousal"
- Probabilidad de acentuación: Habla de lo variable que es el texto. Es la probabilidad de que una sílaba, sin tener un acento gramatical, esté acentuada. Aumenta con "Valence".

Con las siguientes formulas empíricas calculamos los parámetros prosódicos:

$$Volumen = 1 + 0{,}4A + 0{,}4P$$

$$Duracionmedia = 100 - 40A + 25V - 20P$$

$$Variaciondeladuracin = 20 + 20A + 15V - 20P$$

$$Frecuenciamedia = 240{,}3 + 45{,}2A + 20V + 10P$$

$$Variaciondelafrecuencia = 25 + |30A| - |4{,}2V| + |4{,}72P|$$

$$Pendiente = 3A + 2V$$

$$Duraciondelaspausas = duracionmedia/2 - 50A - 50V$$

$$Probabilidaddeacentuacion = V$$

## Recopilación de la información fonética del texto y colocación en array (Función phonems)

Esta función recoge algunos datos provenientes del texto, los estudia y coloca datos que vamos a necesitar en adelante en un vector.

Para tener ordenados los datos que podemos sacar del texto y que van asignados a cada fonema vamos a crear un array de estructuras de este tipo.

```
typedef struct{
char fono;
bool vocal;
bool acented;
bool space;
double dur;
int nword;
}fonema;
```

Con este tipo de estructura crearemos un vector en el que colocaremos la información ordenándola de la siguiente forma.

| H | O | Y | | H | A | Y | | G | U | I | S | A | N | T | E | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

La figura 6.1.2 muestra el array de estructuras ya rellenado con la información de la frase de ejemplo.

| FONOS | - | o | i | - | a | i | - | g | i | s | a | n | t | e | s | - |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| VOCAL | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| ACENTUADA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ESPACIO | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| DURACIÓN | 66 | 111 | 72 | 70 | 59 | 130 | 116 | 73 | 122 | 95 | 102 | 87 | 114 | 99 | 55 | 40 |
| NÚMERO DE PALABRA | 1 | 1 | 1 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 0 |

Figura 6.1.2: Ejemplo de vector de la estructura definida.

Para rellenar el vector lo recorremos a la vez que el que contiene el texto que vamos a sintetizar. Para esto utilizamos dos contadores: “i” para recorrer el texto y “o” para recorrer el vector a rellenar.

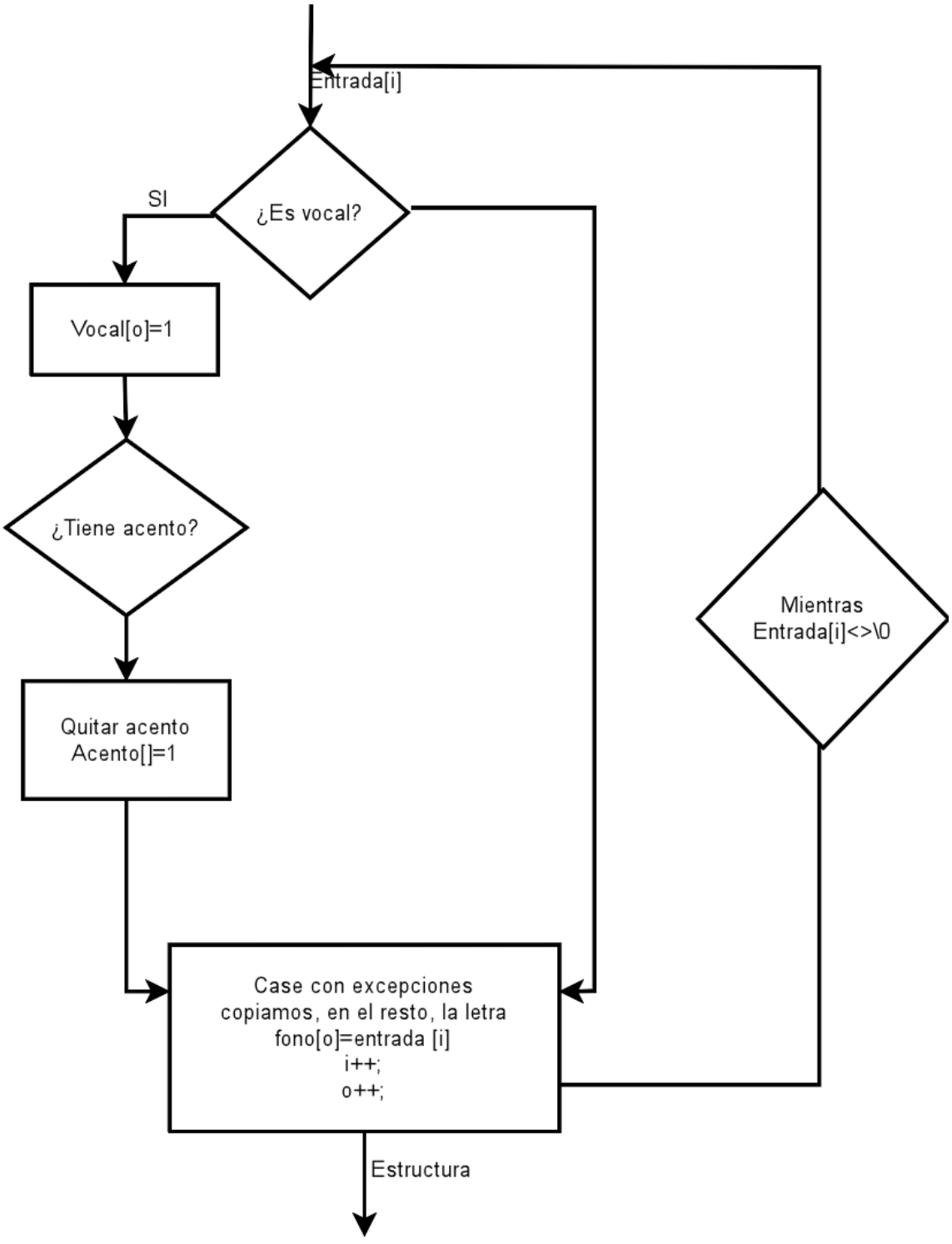



Figura 6.1.3: Rellenado del vector.

Las excepciones de las que hablamos en el gráfico viene dadas por varios motivos, o porque la representación del sonido no es igual a la que reconoce la base de datos, como es el caso de la 'v' que tiene que cambiar a 'b', o porque el sonido dependa de lo que encontremos alrededor como la 'c' o la 'r'.

Los espacios y demás signos de puntuación se cambian por '_' porque es como se repre-

senta el espacio en blanco en nuestra base de datos. Al aparecer el espacio en blanco en el texto, cambiamos la variable “espacio” que corresponda, del vector de salida, a uno (“salida[o].espacio = 1”). La exclamación y la interrogación cambian el modo de la frase, por lo que en el momento en que estos signos son encontrados en el texto, cambia la variable que representa el modo.

Vamos a poner en pseudocódigo por ejemplo la c. “i” es el contador del texto y “o” es el contador de salida.

```
Entra la letra c
Llamamos a quitar acento para entrada[i+1]
Guardamos siguiente letra=entrada[i+1]
Salto al case de 'c'
Swich con la siguiente letra
    si es a
        fono[o]=k
        fono[o+1]=a
        o=o+2
        i=i+1
        Así con el resto de las vocales o en este caso con la h
Vuelve al principio del bucle e inicia el proceso con la siguiente letra.
```

Solo nos falta rellenar numero de palabra (que nos servirá para reconocer la última palabra y matizar el final) y duración.

El número de palabra se determina contando el número de espacios en blanco encontrados en el texto. De forma que todos los fonos entre ese y el siguiente mantendrán ese número de palabra.

Por último tenemos que dar una duración a cada fonema. Para esto usaremos la duración media y la variación de la duración, que hemos calculado en la primera función. A la duración media le sumamos o restamos una fracción aleatoria de la variación de la duración. En lenguaje hablado, el alargamiento de los fonos es percibido como acento. Por ello, la duración de los fonos acentuados se aumentará sumándoles una segunda vez la variación de la duración.

## Selección y adaptación del fichero con la información de la forma de la frecuecia(Fichero buscarfichero y leer pitch.)

En esta sección vamos a elegir la curva principal de la frase. Partiremos de un directorio con varias curvas y de la variable que tiene almacenado el modo de la frase.

En la función “buscarfichero” seleccionamos el nombre del fichero. Los fichero de las curvas están nombrados de esta manera “modo.número”( interrogativa.3). Concatenamos el modo y un numero al azar y tenemos el fichero que contendrá una forma de frase con el modo expresivo que buscamos.

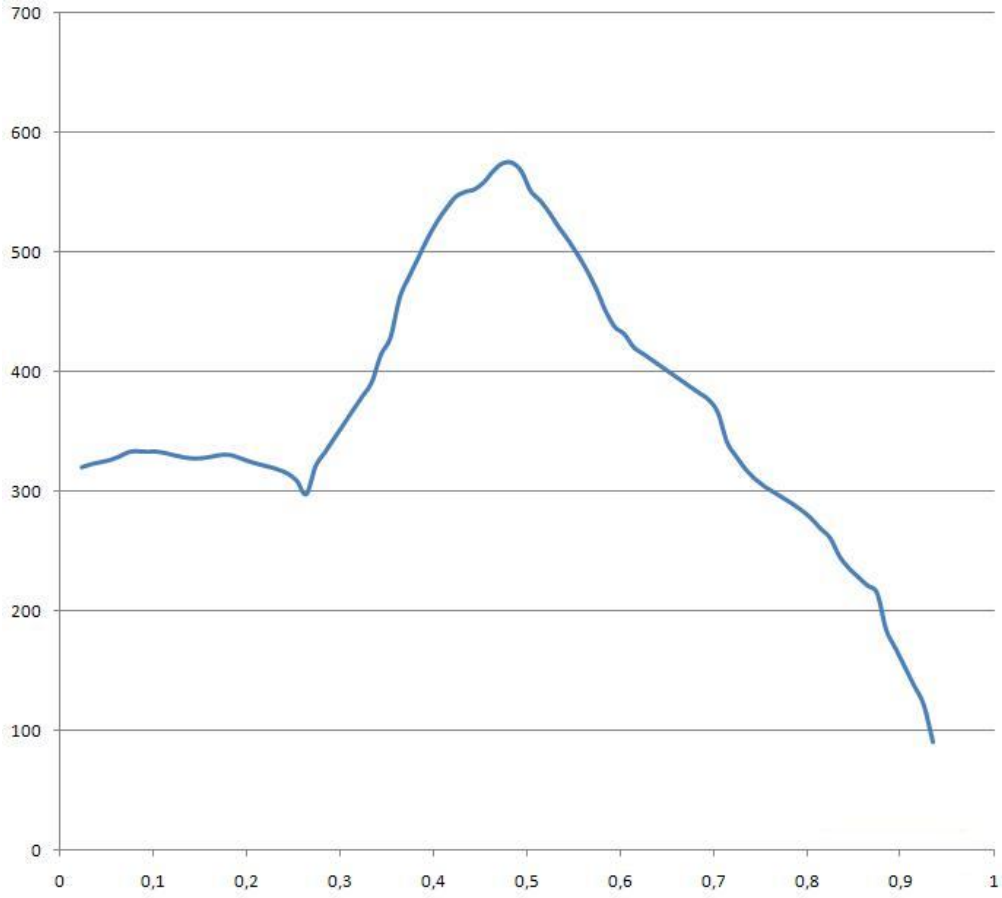



Figura 6.1.4: Ejemplo de forma exclamativa Eje Y:Frecuencia (Hz), EjeX: Tanto por uno del tiempo( ).

Ahora con leerpirch. Abrimos el archivo que hemos “montado” en la función anterior.

Con esos datos vamos a transponer la misma curva a la media que necesitamos y a la duración total que tendrá nuestra frase.

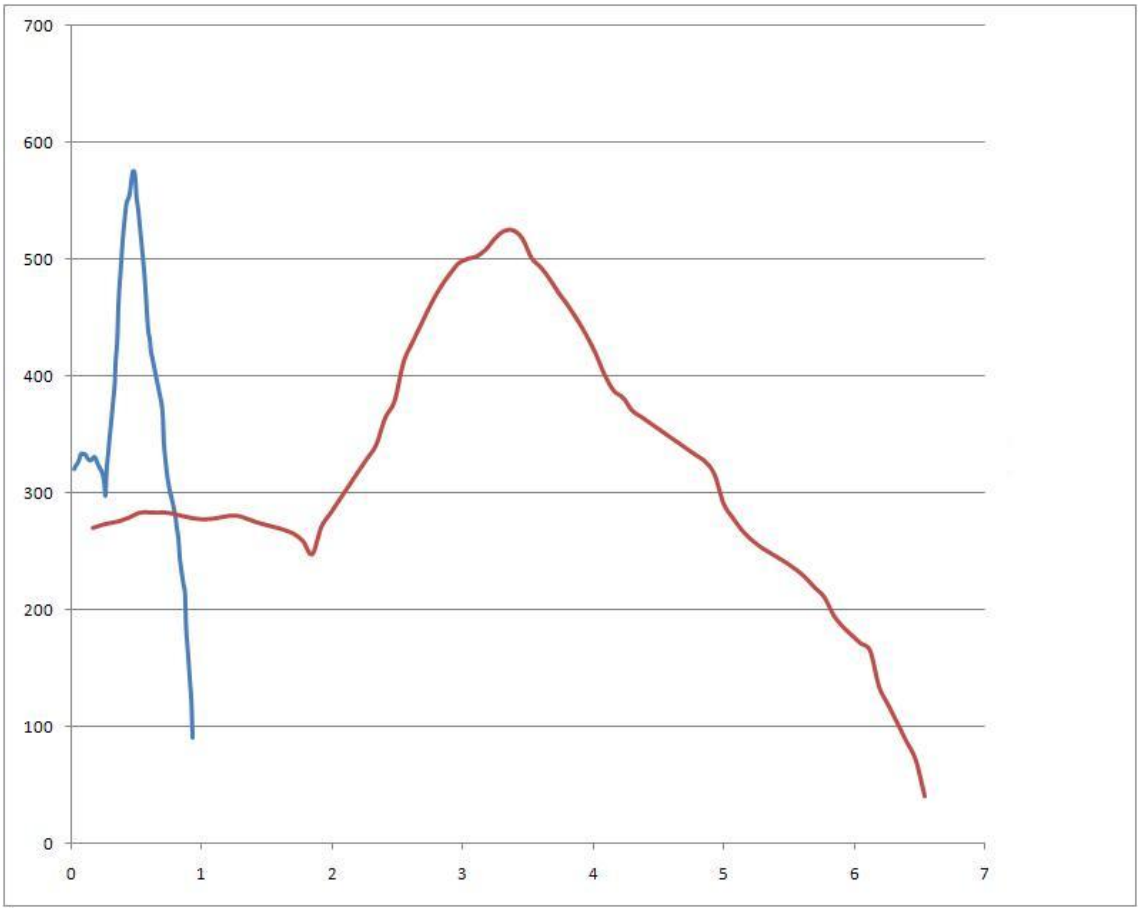



Figura 6.1.5: Adaptación a nuestro sonido. Azul: curva original con duración en [0;1]. Rojo: curva adaptada a la duración del sonido de entrada Eje Y:Frecuecia (Hz); Eje X: Tiempo(s) .

Una vez que tenemos la curva que vamos a seguir vamos a colocar donde van a empezar cada uno de los difonos. Para eso, vamos incrementando el tiempo en el eje X y buscamos cual es el valor en Hz si siguiera la curva anterior.

Este problema de interpolación lo vamos a resolver mediante de la técnica de métodos numéricos Spline. Un spline es una curva diferenciable definida en porciones mediante polinomios.

Vamos a interpolar una función $f(x)$ de la que se nos dan un número N de pares (x,f(x)) por los que tendrá que pasar nuestra función polinómica $P(x)$. Esta serie de funciones nuestras van a ser lineales, esto es, con grado 1: de la forma $P(x) = ax + b$. Definiremos una de estas funciones por cada par de puntos adyacentes, hasta un total de (N-1) funciones, haciéndolas pasar obligatoriamente por los puntos que van a determinarlas, es decir, la función $P(x)$ será el conjunto de segmentos que unen nodos consecutivos; es por ello que nuestra función será continua en dichos puntos, pero no derivable en general.

Obtenemos esta función mediante splines.

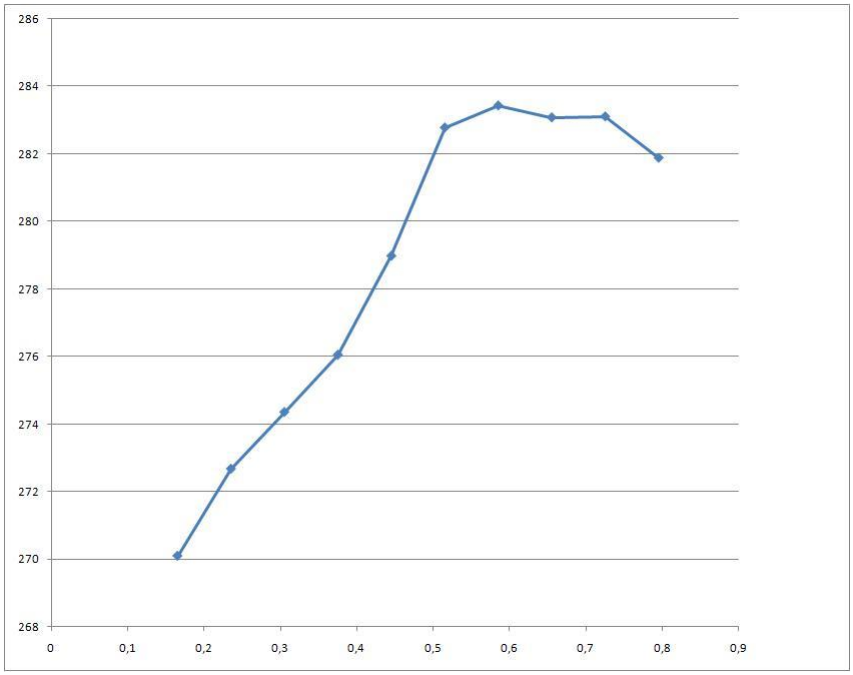



Figura 6.1.6: Curva aproximada. Frecuencia(Hz); Tanto por uno del tiempo.

## Recopilación de toda la información y generación del fichero de entrada a Mbrola(Fichero Crearpho)

Finalmente vamos a asignar la frecuencia concreta a cada fono.

En la primera función hemos adjudicado una duración a cada fono y la hemos guardado en un array con el resto de información sobre el texto. Ahora recuperamos esa duración y la dividimos en fragmentos de 60ms y a cada uno de ellos le vamos a asignar una frecuencia que será la frecuencia asignada por la curva anterior para ese tiempo, sumado, o restado, un fracción aleatoria del parámetro “variación de frecuencia”.

Por ejemplo: a la letra ’o’ en una frase se le ha adjudicado [3]una duración de 136 ms. Ahora asignariamos que tiempo corresponde a la ’o’. En este ejemplo,

| | | |
|---|---|---|
| 1620.2ms | 1680.0ms | 1729ms |

.

Para cada uno de estos tiempos se busca su frecuencia correspondiente en la curva y se le suma o se le resta una fracción aleatoria de la característica “variación de frecuencia” (en este caso es 25Hz).

Un fonema puede estar acentuado por dos razones, porque tenga un acento gramatical o porque se acentúe por la prosodia.[4]Como se indico anteriormente, se le sumará la mitad de la característica “variación de frecuencia” .

[3]en la función phonems, cuando llenamos la estructura.

[4]Cuando estamos alegres o enfadamos ponemos más acentos en las frases, la probabilidad de acento es una caracterítica de las que hemos calculado en el primer apartado.

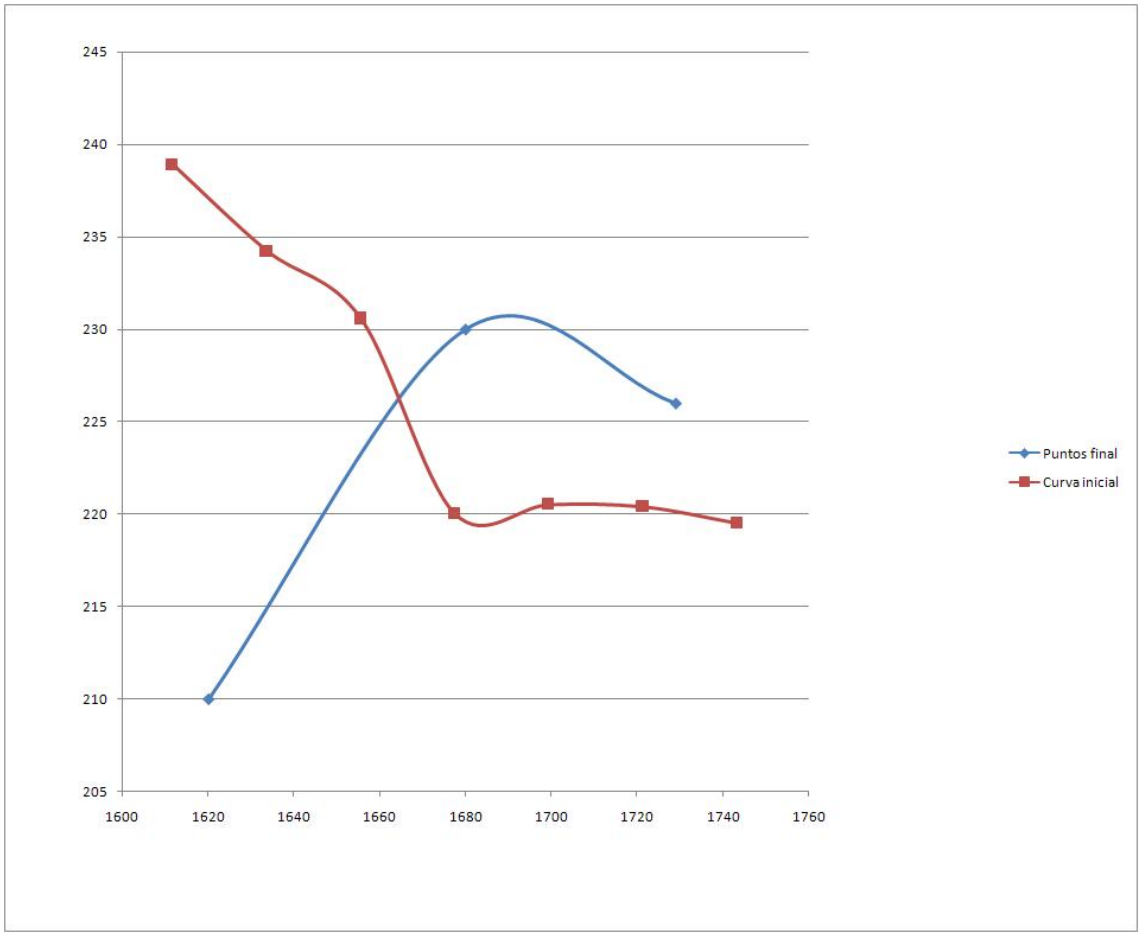



Figura 6.1.7: La 'o' en una de las frases. Eje Y:Frecuencia (Hz); Eje X: Tiempo (ms).

Por último aumentamos o disminuimos la frecuencia de la última palabra para hacer que acabe arriba, por ejemplo cuando estás feliz, o abajo, cuando estás triste. Utilizamos el parámetro de la inclínación que habla de cuanto sube, o baja, la frecuencia por unidad de tiempo. Este parámetro se encuentra entre -3 y 3 y la subirá más rápidamente cuanto más se acerque a tres, o bajara rápido acercándose a menos tres.

Colocaremos los datos en los lugares correspondientes del fichero.

| fono | duración | porcentaje | frecuencia (Hz) | porcentaje | frecuencia(Hz) | ... |
|---|---|---|---|---|---|---|
| e | 110 | 49 | 208 | 100 | 196 | |
| s | 72 | 5 | 203 | 89 | 192 | |
| t | 70 | 74 | 199 | 100 | 242 | |
| ... | ... | ... | ... | ... | ... | |

La señal de voz comparada con la curva de la que partíamos será esta.

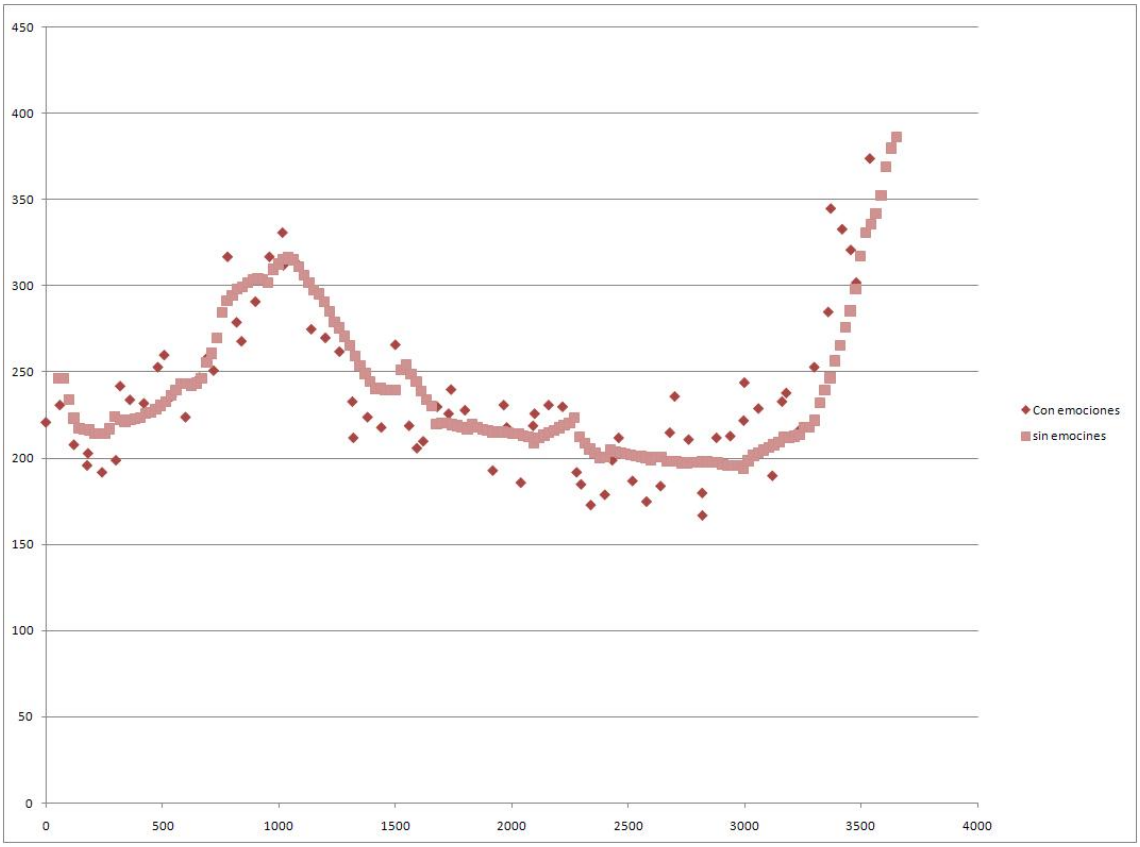



Figura 6.1.8: Puntos de la envolvente con y sin emoción añadida. Eje Y :Frecuencia (Hz) Eje x: Tiempo (ms).

## 6.2. Ejecución de Mbrola, generación del audio y reproducción.

Por último llamaremos a Mbrola para que genere el audio, esto lo haremos en el “main”, completando los parametros de Mbrola con los conseguidos a lo largo de todo el programa. La llamada será:

*sprintf( comando, "%s -e -v%’.3f -f%’.3f -t%’.3f%s%s%s" ,"mbrola", config.volumen, 1.00000, 1.0000, "spanishfemale", pho, argv[6]);*

Esto generará la siguiente línea de ejecución.

*mbrola -e -v <valor del volumen> -f <frecuencia> -t <tiempo> <nombre de la base de datos> <nombre del archivo de extensión .pho> <nombre del archivo de salida>;*

- -e indica que ignorará los errores fatales generados por la falta de algún difono en la base de datos.
- -v espera detrás el valor de volumen, así como -f de frecuencia, -t de tiempo. De estos tres a nosotros solo nos interesa usar el de volumen, ya que, los otros datos

están incluidos en el fichero de entrada a mbrola. Aun así en esta línea podríamos modificarlos en conjunto, es decir, aumentaríamos la frecuencia de cada fono entero o la velocidad de emisión de todos los difonos.

- Nombre de la base de datos de difonos, la cogeremos directamente de la entrada al programa.
- Nombre del archivo de entrada mbrola (.pho). Al crear el archivo nombramos el fichero con el texto que hemos creado sin espacios. Es decir, si vamos a decir "hola mundo" el fichero será "holamundo.pho).
- Fichero de salida. Coge directamente el nombre de los parámetros de entrada del programa.

En linux, el programa abre el reproductor de sonido antes de finalizar su ejecución, para mostrar el resultado

## 6.3. Posibilidad mensaje neutro

Nuestro programa contiene la posibilidad de crear el mensaje sin ningún tipo de emoción. Para esto solo tenemos que ejecutarlo con el primer parámetro en 1.

La consola nos preguntará con que frecuencia queremos el mensaje (Hz) y el tiempo que queremos que dure cada uno (ms).

Con esto creamos directamente el archivo de extensión (".pho").

Finalmente ejecutamos la línea de Mbrola y la del reproductor de audio.

# Capítulo 7

# Evaluación experimental

Como ya hemos vista la percepción de emociones en la voz depende mucho del receptor. Nuestras evaluaciones se tendrán que centrar en cómo llega el mensaje a la población estudiada y ver si, en su mayoría, diferencia entre emociones y percibe el mensaje no solamente por el texto sino también por la prosodia y en qué medida.

Vamos a estudiar estadísticamente, con estas premisas, la calidad del sistema en cuanto a expresión con distintos escenarios [1]

## 7.1. Pruebas realizadas

Para realizar las pruebas haremos un cuestionario on-line en el que haremos distintas preguntas referidas a distintos audios creados con nuestro software. Concluimos entonces que el experimento esta sesgado a la parte de la población que maneja frecuentemente el ordenador y las redes sociales, encontrándose, la mayoría entre los 25 y los 60 años.

El estudio será cualitativo, se basará en la evaluación subjetiva de la calidad del mensaje tanto textual como prosódicamente.

Vamos a someter a experimento los siguientes campos:

- La comprensión textual del mensaje: Cómo se entiende solamente el texto y en que medida esta varía con la variación de la frecuencia y la velocidad.
- La comprensión de la emoción solamente a través de la prosodia comparándola con la misma pero con voz humana.
- La compresión de la emoción en un mensaje que puede tener varios significados.
- La aceptación del mensaje acompañado de la parte prosódica.

[1] Con escenarios nos referimos a otras partes del sistema expresivo que apoyan el significado.

## Comprensión de textual del mensaje

Este experimento mide el grado de inteligibilidad que tienen las frases sin la implementación de la emoción.

Esto lo hacemos para comprobar la calidad de la base de datos de difonos. A los usuarios se les reproducirá una frase y tendran que escribir lo que han entendido y valorar del 1 al 5 la dificultad que han tenido. En las frases se irán incrementando la velocidad y la frecuencia.

Además de puntuar tendrán que escribir el contenido en un cuadro de texto.

**Resultados** La media de la puntuación de los usuarios es de 3.7 (entre 1 y 5).

Si dividimos el experimento en grupos según la velocidad y la frecuencia:

1. La frase con menos velocidad y menos aguda (200Hz) tiene una puntuación de 3,94.
2. Velocidad media a frecuencia media (220Hz) tiene una puntuación de 3,63.
3. Velocidad rápida y frecuencia alta (280Hz) puntúa con un 3,54.

Respecto al contenido, todos los encuestados escriben correctamente el texto, a excepción de algún caso en la versión más rápida que confunde “un dos” con “juntos”.

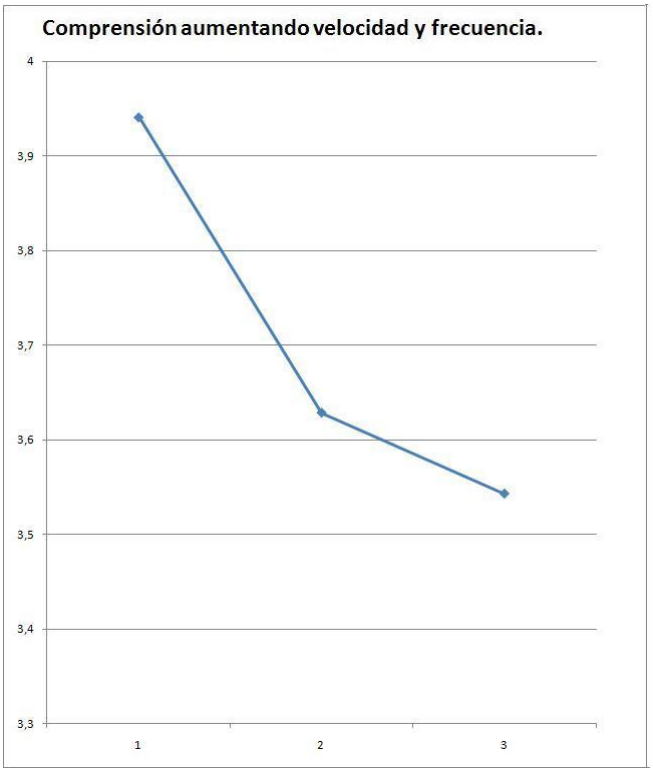



Figura 7.1.1: Influencia de la velocidad y la frecuencia en la inteligibilidad. Eje Y: Media de la puntuacion [1;5]; Eje X: Número del caso.

## Comparación del reconocimiento entre voz sintentizada y voz humana sin apoyo semántico textual

Este experimento mide la capacidad de reconocer la intención solamente con la prosodia y compararemos los resultados obtenidos con los resultados con voces grabadas por humanos.

A los usuarios se les reproducirá una frase sin contenido semántico textual (“lalelilu lilo lilula blabla blible)”y con contenido prosódico y se les pedirá que elijan la emoción que representa entre alegría, tristeza, enfado, miedo, neutro.

Con este experimento vamos a estudiar en que medida la prosodia de la frase por sí sola hace entender la emoción y si la limitación es del software o es una limitación que también se da en la voz humana.

**Resultado:** La matriz de de Confusión para nuestro experimento es la siguiente.

| | miedo | enfado | alegría | tristeza | calma |
|---|---|---|---|---|---|
| miedo | **54** | 14 | 20 | 6 | 6 |
| enfado | 0 | **31** | 20 | 34 | 9 |
| alegría | 3 | 23 | **39** | 0 | 17 |
| tristeza | 34 | 23 | 18 | **40** | 31 |
| calma | 9 | 9 | 3 | 20 | **37** |

Figura 7.1.2: Matriz de confusión propia en %.Las columnas representan la emoción reproducida y las filas la emoción reconocida.

- La elección se representa en el eje de las x con la siguiente leyenda 1-Miedo 2-Enfado 3-Alegría 4-Tristeza 5-Calma.

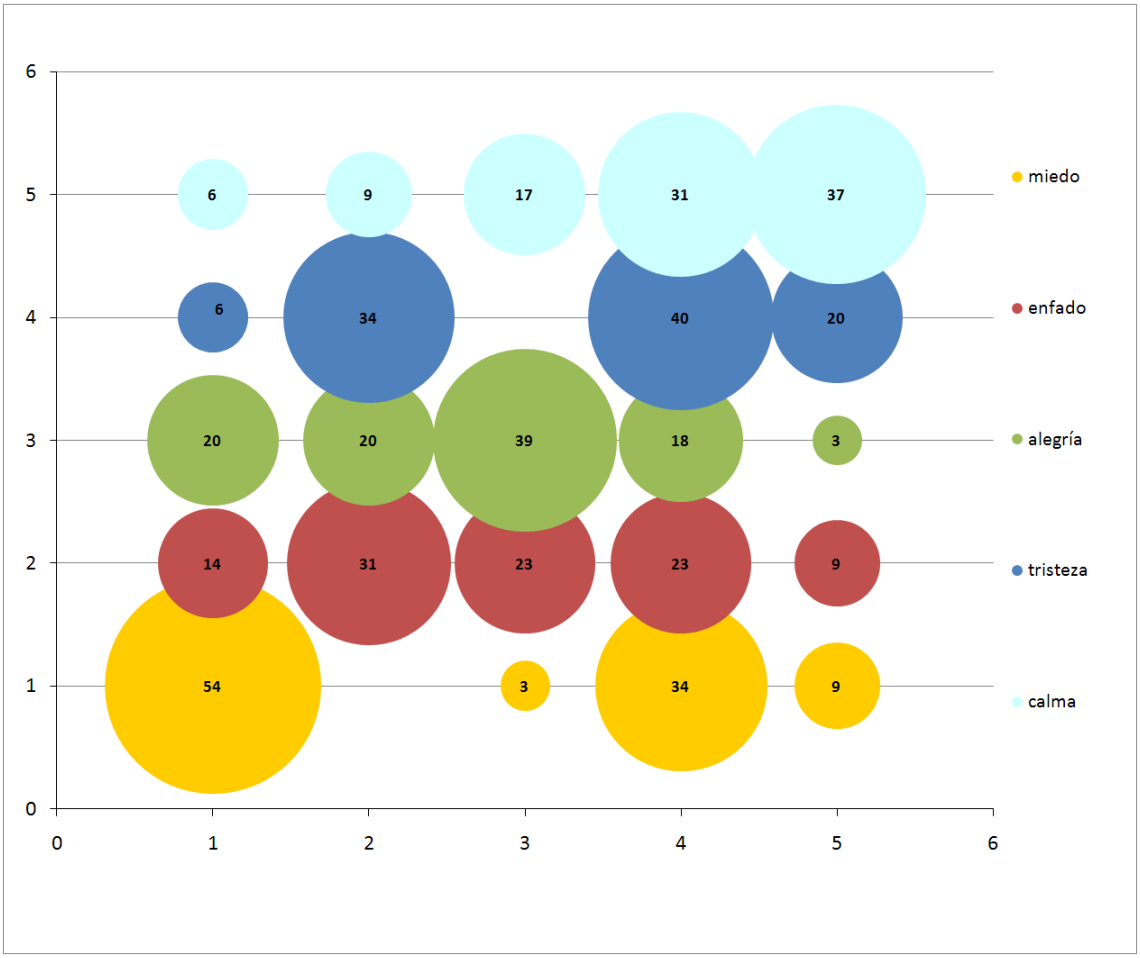



Figura 7.1.3: Comparación de la percepción

- Las emociones más percibidas son las expresadas.
- Aunque, aunque la emoción creadas es la que más se entiende la diferencia con el restod es pequeña.
- La que se diferencia claramente es miedo, aunque se confunde con tristeza.
- Tristeza es la emoción que más se percibe en otras emociones.
- Enfado y alegría tienen una dispersión muy grande en su percepción

## Diferenciación de significado en una frase, que varía solo por la vertiente prosódica, con voz sintetizada.

Ahora vamos a unir las dos cosas, el contenido prosódoco y el textual. Para esto vamos a hacer dos tipos de experimentos:

- Uno en el que vamos a reproducir una frase y medir del 1 al 5 si se adecua al significado
- Otro en el que pondremos el mismo texto y el significado se diferenciará por la prosódia, el usuario tendrá que distinguir el significado de cada uno.

En el primer tipo de experimento pongo 4 frases. “He acabado la universidad”, “Se ha muerto mi abuelo”, “Yo soy tu madre” y “No me hagas nada”. Las puntuaciones medias del 1 al 5 respectivamente son las siguientes.

- “He acabado la universidad” que con una expresión emocional de Arousal, Valence, Power de (1;1;0,75) obtiene de media un 3,9 sobre 5.
- “Se ha muerto mi abuelo” con AVP=(-1;-1;0) obtiene una media de 3,84.
- “Yo soy tu madre” con AVP=(-1,0,1) un 3,43.
- “No me hagas nada” con AVP=(0.5,-1,-1) un 4,19;

En el segundo experimento comparamos que estamos agregando un significado. La frase es “Tenemos ensalada para comer” una de ella se hará con AVP=(0.3;1;0.5) buscando dar a entender que le agrada la idea, y otro con AVP=(-0.3;-1;-0.5) buscando dar a entender que no. Preguntamos a los encuestados si creen que al emisor le gusta la sopa. Del primer caso un 97 % responde que sí y en el segundo caso un 90 % responde que no.

# Capítulo 8

# Conclusiones y líneas futuras de trabajo

## 8.1. Conclusiones

El desarrollo de este proyecto arroja varias conclusiones en función del marco que estudiemos.

#### Comprensión textual del mensaje neutro

La base de datos desarrollada logra el nivel de percepción necesitado. Ya que todos los encuestados entienden el mensaje.

Observamos que la dificultad para entender el mensaje se incrementa con la frecuencia y la velocidad. En nuestro caso, al ser una voz femenina, vamos a encontrar más problemas de inteligibilidad que en una voz masculina.

**Comprensión de la emotividad sin contenido textual.** Del estudio [22] obtenemos la siguiente tabla, para poder comparar nuestros resultados.

**Table 1:** People Performance Confusion Matrix

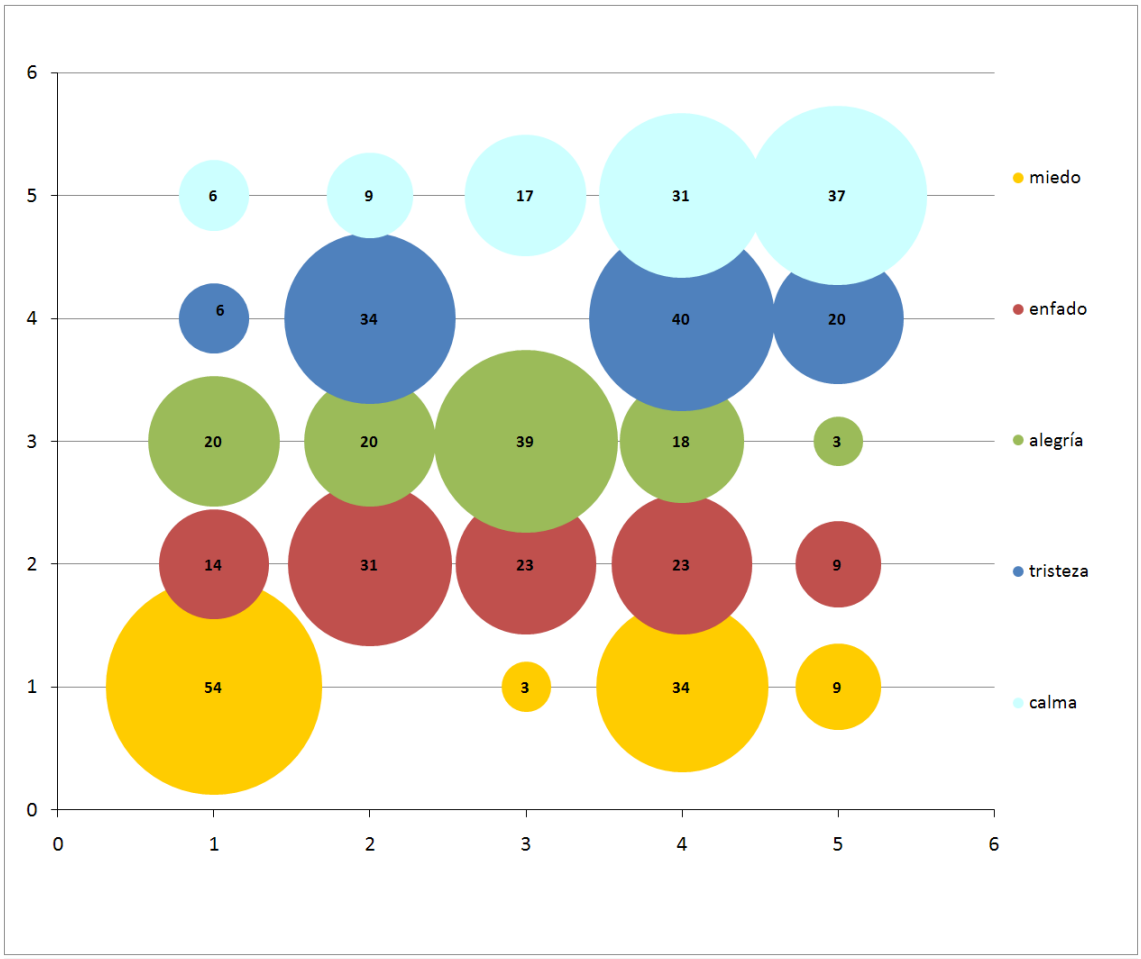
| Category | Normal | Happy | Angry | Sad | Afraid |
|---|---|---|---|---|---|
| **Normal** | **66.3** | 2.5 | 7.0 | 18.2 | 6.0 |
| **Happy** | 11.9 | **61.4** | 10.1 | 4.1 | 12.5 |
| **Angry** | 10.6 | 5.2 | **72.2** | 5.6 | 6.3 |
| **Sad** | 11.8 | 1.0 | 4.7 | **68.3** | 14.3 |
| **Afraid** | 11.8 | 9.4 | 5.1 | 24.2 | **49.5** |

Figura 8.1.1: Matriz de confusión con voces humanas

En la matriz vemos que la comprensión sin la información contextual está muy alejada de ser perfecta. Esto se debe la dependencia de la cultura y del entorno de cada persona, por ejemplo, a todos nos parece que los alemanes están siempre enfadados y los franceses enamorados.

En las siguientes gráficas mostramos los dos estudios comparados.

- La elección se representa en el eje de las x con la siguiente leyenda 1-Miedo 2-Enfado 3-Alegría 4-Tristeza 5-Calma.



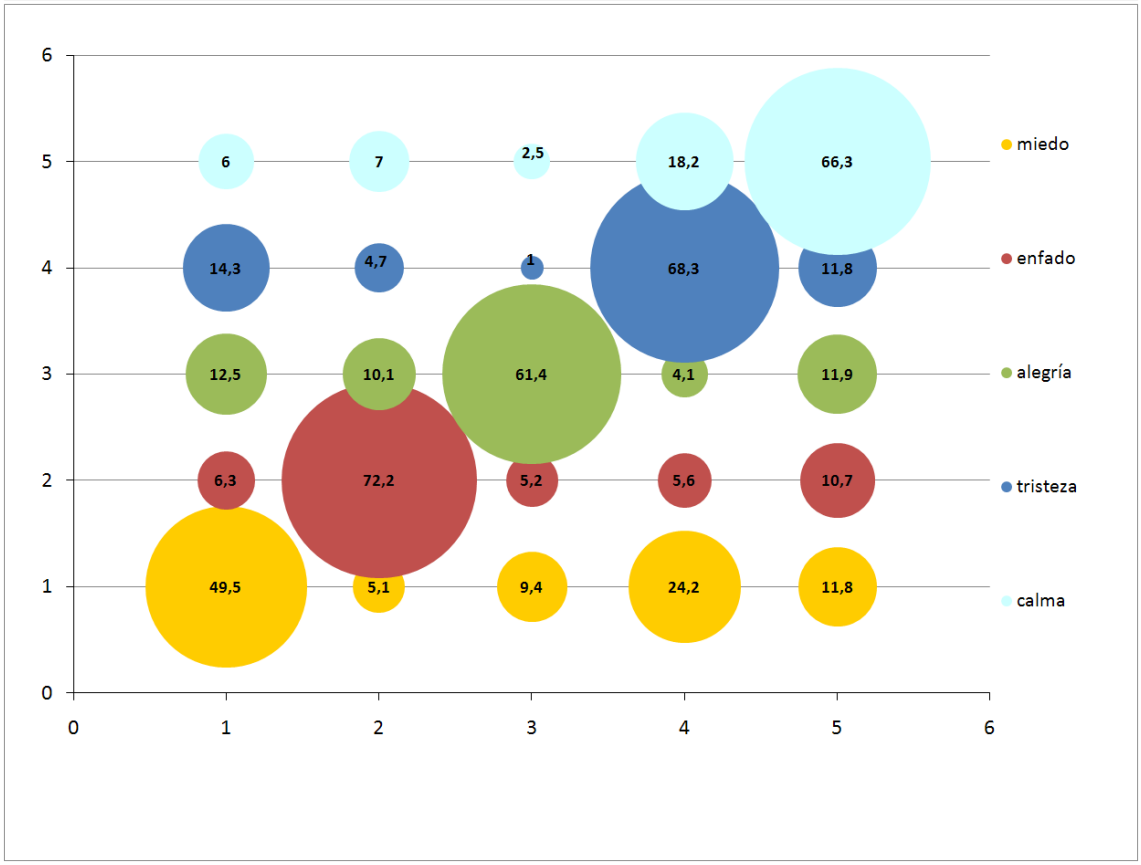




Figura 8.1.2: Comparación de la percepción

Las emociones que mejor se diferencian son miedo y tristeza, esto se debe a que las frases de nuestro sistema se producen lentamente. Esto ha sido necesario por el compromiso entre la inteligibilidad y la comprensión de la emoción. Si las frases se dicen demasiado rápido dejan de entenderse, y por tanto para que se entienda el texto tenemos que bajar la velocidad, en desmedro de la inteligibilidad de las emociones en las que la velocidad del habla es mayor.

Por esta razón las emociones más rápidas se entienden como las lentas (miedo y tristeza).a parte de esta limitación están las limitaciones por la tipología de las propias emociones ya que muchas de ellas tienen una o varias características prosódicas parecidas y al depender solo del resto se nos hace más difícil diferenciarlas. Por ejemplo: Miedo y tristeza se pueden confundir porque la variación de la frecuencia es muy parecida en un texto hablado de forma que la diferenciación termina dependiendo del tono.

Concluyendo, al escuchar el audio, la emoción generada es la más reconocida, pero menos de la mitad de los encuestados reconocen la emoción (excepto para miedo). Podríamos decir que se diferencian pero no de manera muy clara del resto de emociones si no creamos un contexto.

**Diferenciación de emociones con el mismo texto.** En este punto llegamos al objetivo del trabajo. En su mayoría, los usuarios son capaces de inferir información de un mensaje con el mismo texto pero distinta entonación. Se puede conseguir que el mensaje tenga más información que la meramente textual.

Otro de los objetivos del sistema es no trabajar solamente con emociones primarias, en el último experimento, mezclamos las componentes para acompañar al mensaje que estamos dando. No buscamos solo la alegría, sino que queremos alegría con excitación y con el resto de emociones. Las puntuaciones de esta evaluación es alta, los encuestados creen adecuada la prosodia con el texto.

## 8.2. Aportaciones

Este proyecto tiene dos aportaciones importantes, una de ellas es la creación de la base de datos Mbrola y el otro la creación del código usando ese sistema de síntesis que convierte el texto y un punto en un espacio$R^3$ continuo, en un archivo de audio.

- La base de datos colabora con el proyecto Mbrola, dando la posibilidad de su uso en español para una voz femenina.

Desde el punto de vista de la programación, nuestro proyecto aporta.

- Integración de un sistema emotivo en la síntesis de Mbrola en español.
- Sistema de síntesis de voz emotiva con poca necesidad de potencia de proceso y de memoria.
- Un sistema capaz de sacar en tiempo real una voz que esté de acuerdo con el sentimiento de un motor emotivo.

## 8.3. Líneas futuras de trabajo

Durante la elaboración y evaluación del proyecto surgen líneas paralelas y futuras de trabajo.

- Integrar el sistema con un motor emotivo, un generador de texto y un generador de emociones faciales, de manera que la interacción del robot sería emotiva.

### Funcionamiento del programa en un robot

El programa está diseñado para generar y reproducir el sonido hablado correspondiente a un texto y una emoción proporcionados como entrada. El programa es capaz de hacer esto en un espacio de tiempo que permita sostener un ritmo de conversación adecuado.

Para ello las emociones y el texto se recogen de parámetros de entrada del programa. De esta manera será más facil modificar el código para las comunicaciones con los otros procesos que nos generaran el texto y la emoción.

Una posibilidad en el entorno universitario es que el robot esté funcionando con ROS [1](Robot Operating System). Suponiendo que el motor emotivo(nodo) pasa un mensaje (topic) con los datos de la emoción y otro programa(nodo) en ejecución genera un texto que transfiere por otro mensaje (topic). Nuestro programa, que esta definido como nodo, recibirá los datos y se ejecutará creando el sonido.

- Mejora de los parámetros expresivos por aprendizaje con redes de neuronas.
- Sistema de adaptación a cada persona a partir de un sistema de "reset" que grabe al usuario distintas maneras y obtenga de manera subjetiva los parámetros prosódicos.
- Generación de voz cantada.
- Posibilidad de implementar cambios de idiomas y configurar distintos parámetros para varias culturas.

[1]ROS provee los servicios estándar de un sistema operativo tales como abstracción del hardware, control de dispositivos de bajo nivel, implementación de funcionalidad de uso común, paso de mensajes entre procesos y mantenimiento de paquetes.http://es.wikipedia.org/wiki/Sistema_Operativo_Rob%C3%B3tico

# Capítulo 9

# Presupuesto

Los gastos a considerar en esta proyecto son principalmente el tiempo de trabajo de dos trabajadores y el equipo (hardware) utilizado en la grabación y para la programación. El software usado en este proyecto ha sido todo software libre así que no ha generado ningún tipo de gasto.

En el proceso de grabación usamos un micrófono y la tarjeta de sonido del PC.

En cuanto a personal contamos con un ingeniero sénior y un ingeniero durante el año de duración del proyecto (como vemos en la cronología del proyecto, apartado 1.2)

Los costes desglosados quedan de esta manera:

## Coste asociado al personal:

| Apellidos y nombre | Categoría | Dedicación (hombre mes) | Coste hombre mes | coste (euros) |
|---|---|---|---|---|
| Javier Fernández de Gorostiza | Ingeniero Senior | 0,5 | 4289,54 | 2144,77 |
| Emilia Méndez Barrios | Ingeniero | 4,17 | 2694,39 | 11235,6063 |
| | Hombres mes | 4,67 | Total | 13380,3763 |

1 Hombre mes = 131.25 horas. Máximo anual de dedicación de 12 hombres mes (1575 horas)

Máximo anual de PDI de la Universidad Carlos III de Madrid 8.8 hombre mes (1155 horas)

## Coste asociado a los equipos:

| Descripción | Coste (Euros) | % Uso dedicado al proyecto | Dedicación (meses) | Periodo de depreciación | Coste imputable (euros) |
|---|---|---|---|---|---|
| Ordenador | 1000 | 100 | 12 | 60 | 100 |
| Micrófono | 90 | 100 | 1 | 60 | 1,5 |
| Alimentación Phantom | 45 | 100 | 1 | 60 | 0,75 |
| | | | | Total | 102,25 |

La formula del calculo de la Amortización es la siguiente:

Número de meses, desde la fecha de facturación. que el equipo es utilizado · Coste del equipo( sin IVA) ·% del usado que se dedica al proyecto / Periodo de depreciación

Por tanto el coste total del proyecto es de 13482,63€

# Bibliografía


[1] Nicolas D Alessandro. The MBROLA Synthesizer. 1993.

[2] David L Bimler and Galina V Paramei. Facial-expression affective attributes and their configural correlates: components and categories. *The Spanish journal of psychology*, 9(1):19–31, May 2006.

[3] Janet E Cahn. The Generation of Affect in Synthesized Speech. pages 1–19.

[4] Mariano Cholíz. Psicología de la emoción: El proceso emocional. 2005.

[5] Humberto Pérez Espinosa, Carlos Alberto, Reyes García, Reporte Técnico No Ccc, Luis Enrique Erro, and Sta Ma Tonantzintla. Reconocimiento de Emociones a Partir de Voz Basado en un Modelo Emocional Continuo. (Ccc), 2010.

[6] Isabel Fernández. Bla,bla,bla Hablar, habla cualquera. pages 1–28.

[7] Agustín Gravano. Síntesis Concatenativa. 2012.

[8] P Ji and P Anna. Spectral Properties and Prosodic Parameters of Emotional Speech in Czech and Slovak. (Scherer 2003), 2007.

[9] Tom Johnstone and Klaus R Scherer. The effects of emotions on voice quality.

[10] Hisao Kuwabara. Acoustic characterisitics of speaker individuality : Control and conversion. 16:165–173, 1995.

[11] Jens Madsen. Modeling of Emotions expressed in Music using Audio features. 2011.

[12] Ana María, D Ph, and Michele Dufey. Expresión y reconocimiento de emociones : un punto de encuentro entre evolución , psicofisiología y neurociencias. 2:8–20, 2007.

[13] Mimmi Forsell Master and Science Thesis Stockholm. Acoustic Correlates of Perceived Emotions in Speech Acoustic Correlates of Perceived Emotions in Speech. 2007.

[14] McGraw-Hill. La comunicación no verbal.

[15] Albert Mehrabian. Nonverbal Communication.

[16] Sylvie Mozziconacci. Prosody and Emotions.

[17] Iain R. Murray and John L. Arnott. Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech. *Computer Speech & Language*, 22(2):107–129, April 2008.

[18] Angel Olaz. Diseño de un sistema codificado de notación paralingüística para entrevistas cualitativa. pages 1–17.

[19] Pierre-yves Oudeyer. The Synthesis of Cartoon Emotional Speech.

[20] Mario De Oyarbide. Moduladores de sonido y sus aplicaciones.

[21] Cécile Pereira. Dimensions of emotional meaning in speech. pages 1–4.

[22] Valery A Petrushin. EMOTION RECOGNITION IN SPEECH SIGNAL : EXPERIMENTAL STUDY , DEVELOPMENT , AND APPLICATION. (Icslp), 2000.

[23] Oudeyer Pierre-Yves. The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, 59(1-2):157–183, July 2003.

[24] Erhard Rank and Hannes Pirker. GENERATING EMOTIONAL SPEECH WITH A CONCATENATIVE. pages 1–4.

[25] Darcy Rojas. Prosodia y emociones : datos acústicos , velocidad de habla y percepción de un corpus actuado. (1):59–72.

[26] Marc Schröder. Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions.

[27] Marc Schröder, Roddy Cowie, Ellen Douglas-cowie, Machiel Westerdijk, and Stan Gielen. Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis. pages 1–4.

[28] Raquel Tato, Rocío Santos, Ralf Kompe, and J M Pardo. Sony International ( Europe ) GmbH Universidad Politécnica de Madrid. pages 1–4.

[29] Zhongzhe Xiao, Emmanuel Dellandrea, Weibei Dou, and Liming Chen. A Dimensional Emotion Model Driven Multi-stage Classification of Emotional Speech. pages 1–36.

# ANEXOS

# ANEXO 1 Frases y difonos sacados de cada palabra.

El objetivo es admitir la obsesión. /objetivo ob bj /admitir dm mi /obsesion bs se es si .

El edil eficaz lee con esfuerzo. /edil di il /eficaz ef fi / lee ee e- /esfuerzo sf rz zo.

El pliegue del frijol causa un conflicto ecológico. /pliegue eg ge /frijol -f fr ij jo ol /causa ca au us /conflicto nf fl /ecológico ek ko oj ji.

La doncellita adquirió una teja. /doncellita ey yi /adquirio dk /teja ej ja.

El señor Punset obtuvo ceñudo el correo subrayado / punset pu ns t- /obtuvo bt tu bo /ceñudo eñ ñu do / correo oR Re eo / subrayado –s su br ay.

La epifanía aglutina alforjas de naftalina /epifanía ep pi if fa/ aglutina gl ut / alforjas lf or rj / naftalina -n na ft ta.

La espada de esgrima tiene un cachito corchudo /esgrima sg gr im / cachito ach chi / corchudo rch chu .

Ser esbirro conlleva ser adlátere. /esbirro sb bi iR / conlleva ny ye eb ba / adlátere ad.

Con encanto urdió el jersey. /encanto nk / jersey rs. // en la union de “encato urdió” encontramo el difono ou

Esnifó disyuntivamente ahora. /Esnifo sn ni /disyuntivamente sy / ahora ao.

La guitarra entró en acción en la aljaba. / guitarra gi Ra / acción kz io n- / aljaba lj ab.

Enun lugar de la Mancha hay mucho hidalgo omnipotente. /Enun –e en un / lugar lu ug ga / mancha –m cha / mucho mu uch cho / hidalgo id da al lg go / hay ai /omnipotente om mn ot.

Ruiz toca arte armonico. /Ruiz Ru ui iz z- / toca ok / arte rt / armonico rk.

El hippie pelma fue aclamado por el altavoz de la ciudad. /hippiip i- / pelma pe em /aclamado cl/ altavoz lt oz / ciudad iu d-.

En elche hacen calzado con pelvis de alcón. /elche lch / calzado lz / pelvis lb is / alcon lc.

En el referéndum los búhos se veían impotentes /referéndum –R er um m- / búho -b uo.

El pich es chungo. /pichich ch- / chungo un ng.

El adverbio oficialmente es conmutativo /adverbio db rb bi / oficialmente of / conmutativo nm.

La cuñita de la canoa tiene la culpa /cuñita uñ ñi /canoa –k no oa / culpa lp.

Enrique tuvo una odisea con ogros en Samsara. /Enrique nR Ri ke / odisea od / ogros og os ro / samsara ms.

En el solsticio Esperanza se pone susceptible a los erpes. /solsticio so ls / esperanza sp nz za a-/ susceptible sz pt / erpes rp.

El gasoil corroe los hornos con burbujas de escape. /gasoil oi /corroe oe Ro / hornos rn /burbuja ur uj /escape sk .

El kayak de bambu tiene ocho metros de eslora. / kayak –k/ bambu mb m- / ocho och / eslora sl / tiene -t ti ie en ne.

Apresuradamente paré en el stop./ apresuradamente pr –a re ad me e- / stop op p- st/ Irlanda rl la an.

El hoyuelo por osmosis. Hoyuelo oy/ osmosis sm mo s- .

Ipsofacto dio el argumento de la estufa y el caldo. /ipsofacto ps ak /argumento rg /estufa uf /caldo ld.

El huérfano poco cognitivo tiene una cruz un atlas y una orca /huérfano rf /cognitivo gn / cruz uz ru/ atlas tl /orca rk .

Toco el saxo en el futbol. /saxo ks / futbol tb bo.

El dueto trágico obligó a decir ñoñerías / dueto –d du ue et to / trágico tr ra aj ji ik/ obligo –o obbl li ig / decir –d de ez zi ir / ñoñerías –ñ ño oñ ñe ri s-.

El lleva gañan el cuádruple de tiempo. /gañan ga an / cuádruple –k ku ua

En el cerro llueve leche. / Cerro –z Ro o- / llueve yu -y be / leche ech che.

Dale a Chema agua para que se afeite. / Chema –ch em ma / agua ag gu / afeite af fe ei it te.

Treame un kebab rico de abdul. / tráeme ae /kebab b- / abdul bd ul .

Iñaki incho la rueda a lo burro. / Iñaki iñ ki / incho nch / burro bu uR.

Amparo se endeuda al canjear el azúcar. /Amparo am mp pr o- /endeuda nd en ud en da / canjear an nj ea je ka r- / azúcar az zu uk.

Juan ponte la zapatilla izquierda. / juan ju ua _j / ponte po on nt te / zapatilla –z za ap pa at iy ya/ izquierda rd zk -i

El tsunami sin palabras evito el dogma. / tsunami ts /palabras br / dogma gm.

Cou tiene tecnología. /cou ou / tecnología kn.

Cuyo caso /cuyo uy yo.

Frases con dífonos que solo son entre palabras se hacen con la unión de las palabras: “soñad reloj árbol referendum ningún stop castor coches maíz” Con las palabras “bonito colorido difícil grande chato jorobado cálido largo menudo negro ñoño pálido Raro suelto teñido lleno ceñido”.

# ANEXO 2: Fichero. pho

_ 66 0 221 91 231

e 111 49 208 100 196

s 72 5 203 89 192

t 70 74 199 100 242

e 59 71 234

_ 130 33 232 79 253 100 260

e 116 28 236 79 224

s 73 49 247 100 259

_ 122 19 251 68 317 100 279

u 95 22 268 85 291

n 102 45 317 100 331

_ 87 4 312 73 313

e 114 32 275 85 270

j 99 44 262 100 233

e 55 8 212

m 58 17 224

p 86 13 218 83 266 l 78 57 219 100 206

o 136 20 210 64 230 100 226

_ 89 12 240 79 228

p 148 28 218 68 193 100 231

a 126 10 218 58 186 100 219

r 136 5 226 49 231 93 230

a 70 73 192 100 185

_ 134 31 173 76 179 100 199 l 110 25 212 79 187

a 137 27 175 71 184 100 215

_ 139 15 236 58 211 100 180

m 50 2 167

e 128 9 212 55 213 100 222

m 99 3 244 64 229

o 66 37 190 100 233

r 89 20 238 88 217

i 120 41 253 91 285 100 345

a 86 57 333 100 321

_ 195 12 302 42 374

# Anexo 3: Encuesta realizada

## Encuesta sobre resultados de desarrollo de síntesis de voz emotiva(http://emilia.pusku.com/)

Responde las siguiente pregunta sobre los archivos de audio que encontrarás en http://emilia.pusku.com/, responde en esta pagina no en la de los audios.No te comas la cabeza, lo que primero percibas será el mejor resultado. Muchísimas gracias por vuestra colaboración.

*Obligatorio

### ¿Entiendes la frase?

Escribe lo que dice y puntua del 5(si te ha sido facil entenderla) al 1 si no.

1. **audio numero 1.**
   Escribe lo que dice y puntua del 5(si te ha sido facil entenderla) al 1 si no.
   *Marca solo un óvalo.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| dificil de entender | ◯ | ◯ | ◯ | ◯ | ◯ | fácil de entender |

2. _____

3. **audio numero 2.**
   Escribe lo que dice y puntua del 5(si te ha sido facil entenderla) al 1 si no.
   *Marca solo un óvalo.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| difícil de entender | ◯ | ◯ | ◯ | ◯ | ◯ | facil de entender |

4. _____

5. **audio numero 3.**
   Escribe lo que dice y puntua del 5(si te ha sido facil entenderla) al 1 si no.
   *Marca solo un óvalo.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| difícil de entender | ◯ | ◯ | ◯ | ◯ | ◯ | fácil de entender |

6. _____

Figura 9.0.1: Encuesta realizada. Pagina 1.

# Encuesta sobre resultados de desarrollo de síntesis de voz emotiva(http://emilia.pusku.com/)

Responde las siguiente pregunta sobre los archivos de audio que encontrarás en http://emilia.pusku.com/, responde en esta pagina no en la de los audios.No te comas la cabeza, lo que primero percibas será el mejor resultado.  Muchísimas gracias por vuestra colaboración.

*Obligatorio

## ¿Entiendes la frase?

Escribe lo que dice y puntua del 5(si te ha sido facil entenderla) al 1 si no.

1. **audio numero 1.**
   Escribe lo que dice y puntua del 5(si te ha sido facil entenderla) al 1 si no.
   *Marca solo un óvalo.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| dificil de entender | ◯ | ◯ | ◯ | ◯ | ◯ | fácil de entender |

2. _____

3. **audio numero 2.**
   Escribe lo que dice y puntua del 5(si te ha sido facil entenderla) al 1 si no.
   *Marca solo un óvalo.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| difícil de entender | ◯ | ◯ | ◯ | ◯ | ◯ | facil de entender |

4. _____

5. **audio numero 3.**
   Escribe lo que dice y puntua del 5(si te ha sido facil entenderla) al 1 si no.
   *Marca solo un óvalo.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| difícil de entender | ◯ | ◯ | ◯ | ◯ | ◯ | fácil de entender |

6. _____

Figura 9.0.2: Encuesta realizada. Página 1.

Escucha el audio indicado en la página http://emilia.pusku.com/ y contesta, en esta página, a que emoción de recuerda.

7. **audio numero 4** *
   *Marca solo un óvalo.*

   - ◯ Alegría
   - ◯ Miedo
   - ◯ Calma
   - ◯ Enfado
   - ◯ Tristeza

8. **audio numero 5** *
   *Marca solo un óvalo.*

   - ◯ Alegría
   - ◯ Miedo
   - ◯ Calma
   - ◯ Enfado
   - ◯ Tristeza

9. **audio numero 6** *
   *Marca solo un óvalo.*

   - ◯ Alegría
   - ◯ Miedo
   - ◯ Calma
   - ◯ Enfado
   - ◯ Tristeza

10. **audio numero 7** *
    *Marca solo un óvalo.*

    - ◯ Alegría
    - ◯ Miedo
    - ◯ Calma
    - ◯ Enfado
    - ◯ Tristeza

11. **audio numero 8** *
    *Marca solo un óvalo.*

    - ◯ Alegría
    - ◯ Miedo
    - ◯ Calma
    - ◯ Enfado
    - ◯ Tristeza

## ¿Les gusta la ensalada?

Contesta sí o no según creas.

12. **audio numero 9** *
    *Marca solo un óvalo.*

    - Si
    - No

13. **audio numero 10** *
    *Marca solo un óvalo.*

    - Si
    - No

## ¿Te suena coherente esta frase?

¿Te suena bien la frase? Evalúa del 1 al 5 cuanto se adecúa la entonación al contenido.

14. **audio numero 11**
    *Marca solo un óvalo.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| no se adecúa nada | ◯ | ◯ | ◯ | ◯ | ◯ | totalmente adecuado |

15. **audio numero 12**
    *Marca solo un óvalo.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| no se adecúa nada | ◯ | ◯ | ◯ | ◯ | ◯ | totalmente adecuado |

16. **audio numero 13**
    *Marca solo un óvalo.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| no se adecúa nada | ◯ | ◯ | ◯ | ◯ | ◯ | totalmente adecuado |

17. **audio numero 14**
    *Marca solo un óvalo.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| no se adecúa nada | ◯ | ◯ | ◯ | ◯ | ◯ | totalmente adecuado |

Figura 9.0.4: Encuesta realizada. Página 3.

18. **El último sonido es una fricada de regalo!**

Con la tecnología de
Google Drive

Figura 9.0.5: Encuesta realizada. Página 4.

# Anexo 4: Muestra de los audios


Figura 9.0.6: Encuesta realizada. Audios.